

# ASYMPTOTIC FREQUENTIST COVERAGE PROPERTIES OF BAYESIAN CREDIBLE SETS FOR SIEVE PRIORS IN GENERAL SETTINGS

BY JUDITH ROUSSEAU<sup>\*,†,§</sup>, BOTOND SZABO<sup>†,¶,||</sup>

*University Paris Dauphine<sup>‡</sup>, and CREST-ENSAE<sup>§</sup>, and Budapest  
University of Technology<sup>¶</sup>, and Leiden University<sup>||</sup>*

We investigate the frequentist coverage properties of Bayesian credible sets in a general, adaptive, nonparametric framework. It is well known that the construction of adaptive and honest confidence sets is not possible in general. To overcome this problem we introduce an extra assumption on the functional parameters, the so called “general polished tail” condition. We then show that under standard assumptions both the hierarchical and empirical Bayes methods give adaptive and honest confidence sets for sieve type of priors in general settings. We apply the derived abstract results to various examples, including the nonparametric regression model, density estimation using exponential families of priors, density estimation using histogram priors and nonparametric classification model.

**1. Introduction.** Uncertainty quantification is of key importance in statistical sciences. Estimators without proper uncertainty quantification have only limited practical applicability, since they contain only limited amount of information about their accuracy. In statistics uncertainty about an estimator is described with the help of confidence sets. Confidence statements are then widely used in statistical practice for instance in hypothesis testing. The construction of confidence sets can be however very challenging, especially in complex, nonparametric problems.

A very popular aspect of the Bayesian approach is the built-in, straightforward way of quantifying uncertainty. Especially in high-dimensional and nonparametric problems the posterior distribution is visualized with the help of credible sets, i.e. sets with prescribed (typically 95%) posterior probability. By accumulating large fraction of the posterior mass these sets describe

---

<sup>\*</sup>The project was partially supported by the ANR IPANEMA, the labex ECODEC

<sup>†</sup>The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637. Research partially supported by Netherlands Organization for Scientific Research NWO

*AMS 2000 subject classifications:* Primary 62G20, 62G05; secondary 62G08, 62G07

*Keywords and phrases:* Uncertainty quantification, coverage, posterior contraction rates, adaptation, empirical Bayes, hierarchical Bayes, nonparametric regression, density estimation, classification, sieve prior

the remaining uncertainty of the Bayesian procedure. Due to the existing computational machinery of Bayesian techniques (eg. MCMC, ABC,... etc) these sets are widely used in practice for uncertainty quantification. However, only little is known about their theoretical properties. In parametric models following the celebrated Bernstein-von Mises theorem, credible sets are asymptotically confidence sets as well, laying the base of the practical applicability of the Bayesian approach in simple models.

However, in nonparametric models the question is still unanswered about how much we can trust Bayesian credible sets as a measure of confidence in the statistical procedure from a frequentist perspective. The first results in the nonparametric paradigm were discouraging, showing that the Bernstein-von-Mises theorem does not hold in general, i.e. even in the standard Gaussian white noise model using conjugate Gaussian priors the resulting credible sets have frequentist coverage tending to one, see [15, 16].

Since then the investigation of frequentist coverage properties of Bayesian credible sets have attracted a lot of attention in nonparametric problems. Various approaches were proposed to solve this problem. In [24, 49] the authors verified that by slightly undersmoothing the prior one can still achieve credible sets with good frequentist coverage and minimax size in the same setup as [15]. Another possibility is to consider weaker, negative Holder-norms and derive the Bernstein-von-Mises theorem in the corresponding Banach-space, see [11, 12, 25].

The preceding results are all based on the knowledge of the regularity of the true underlying function, which is in practice generally not available. A more challenging problem is the construction of Bayesian based confidence sets in the adaptive setting where no information is available on the smoothness of the truth. This, however, turns out to be too much to ask for. In [7, 8, 26, 35] it was shown that it is impossible to construct adaptive confidence sets in general.

More precisely assume that the true (functional) parameter  $\theta_0$  belongs to some regularity or sparsity class  $\Theta^\beta$ , indexed by some (unknown) hyperparameter  $\beta$  belonging to some set  $B$ . When  $\beta$  is unknown, the confidence set  $\hat{C}$  cannot depend on it and it is said to be optimal adaptive if first it has uniform coverage:

$$(1) \quad \liminf_n \inf_{\theta_0 \in \cup_{\beta \in B} \Theta^\beta} P_{\theta_0}^{(n)}(\theta_0 \in \hat{C}) \geq 1 - \alpha$$

and second its size is optimal within each parameter class  $\Theta^\beta$ , i.e. for some

$K > 0$

$$(2) \quad \liminf_n \inf_{\beta \in B} \inf_{\theta_0 \in \Theta^\beta} P_{\theta_0} \left( \sup_{\theta_1, \theta_2 \in \hat{C}} d(\theta_1, \theta_2) \leq Kr_{n,\beta} \right) \rightarrow 1,$$

where  $r_{n,\beta}$  is the minimax estimation rate within the class  $\Theta^\beta$ .

As mentioned earlier it is impossible to satisfy both the coverage and the minimax size requirement on the confidence sets in general. To solve this problem additional assumptions were introduced on the parameter value  $\theta_0$  making the construction of adaptive confidence sets possible by discarding certain inconvenient parameters  $\theta_0$ . A frequently applied assumption is self-similarity where it is assumed that the true parameter has similar “local” and “global” behaviour, see for instance [6, 14, 19, 27, 30, 41]. Another approach is to discard parameters which make it impossible to test between the classes  $\Theta^\beta$ . This approach was considered in various models in context of regularity classes in [7, 9, 21] and in sparse high dimensional models [10, 28].

It is a known fact that Bayesian credible balls associated to posterior distribution which concentrate at the minimax rate verify (2), see [22]. The question is then to understand their frequentist coverage and in particular to characterize subsets of  $\cup_\beta \Theta^\beta$  over which (1) is verified as well.

In [44] the authors have generalized the self-similarity assumption introducing the so called polished tail assumption, discussed in this article also in more details. The polished tail (and self-similarity) assumption was then applied in nonparametric regression with rescaled Brownian motion prior [40] and spline priors [39] and in the context of Gaussian white noise model with Gaussian priors constructing  $L_\infty$ -credible sets [42]. Furthermore, an adaptive version of the nonparametric Bernstein von Mises theorem was given in context of the Gaussian white noise model using conjugate Gaussian priors and spike-and-slab prior [32] under the self-similarity assumption. The polished tail assumption was then slightly extended by the implicit extended bias assumption introduced in context of the Gaussian white noise model [2] and applied in sparse high dimensional models with empirical Bayes Gaussian priors [3] and with hierarchical and empirical horseshoe prior [46]. Besides, discarding parameters making testing between classes impossible was considered in [43] also in the context of the Gaussian white noise model.

All the above mentioned papers consider specific choices of the model and the prior distribution and use explicit, conjugate computations which obviously have their limitations. Although these papers already shed lights on certain aspects of Bayesian uncertainty quantification, they do not provide a clear understanding of the underlying general phenomena. A general approach for understanding the coverage of credible sets is still missing. Besides for many nonparametric models and priors no conjugate computation

is possible and therefore they can not be handled using direct computations. In this work we aim to fill this gap and contribute to the fundamental understanding of this rapidly growing field. We derive abstract results under mild conditions for general choices of models and prior distributions, in the spirit of [17, 18, 36].

*1.1. Setup and Notations.* We consider observations  $\mathbf{Y} \in \mathcal{Y}$  distributed from  $P_\theta^{(n)}$ ,  $\theta \in \Theta$ , which are absolutely continuous with respect to a given measure  $\mu$  with density  $p_\theta^{(n)}$ . We denote by  $\ell_n(\theta) = \log p_\theta^{(n)}$  the log-likelihood and throughout the paper  $\theta_0$  designates the true value of the parameter. We denote by  $E_\theta^{(n)}$  expectation with respect to  $P_\theta^{(n)}$ .

In our analysis we consider models defined by

$$(3) \quad \Theta = \cup_{k \in \mathcal{K}} \Theta(k), \quad \Theta(k) \subset \mathbb{R}^{d_k}, \quad d_k \uparrow \infty$$

with  $d_k \asymp k$  and  $\mathcal{K} \subset \mathbb{N}$ . These models are very popular and frequently used in practice, see for instance [20, 45] for a review.

The parameter  $k$  drives the sparsity or the regularity of the model. Finding the model  $\Theta(k)$ , which is the most appropriate for recovering  $\theta_0$ , requires additional information about the true parameter (e.g. regularity, sparsity, ... etc) which is usually not available. Therefore a natural approach is to let the data decide about the optimal model  $\Theta(k)$ . In the Bayesian framework one can accomplish this by the hierarchical or the empirical Bayes approach. In the hierarchical (or also referred to as full Bayes approach) one endows the hyperparameter  $k$  with a prior distribution  $\pi_k$  and conditionally on  $k$ , considers a prior distribution  $\pi_{|k}$  on  $\theta \in \Theta(k)$ , resulting in a prior distribution  $\pi$  on  $\Theta$  defined by:

$$(4) \quad k \sim \pi_k, \quad [\theta|k] \sim \pi_{|k}.$$

We denote the posterior distribution on  $\Theta$  by  $\pi(\theta|\mathbf{Y})$  and the conditional distribution of  $\theta|\mathbf{Y}, k$  by  $\pi_{|k}(\theta|\mathbf{Y})$ .

In contrast to this in the empirical Bayes approach one constructs a frequentist estimator  $\hat{k}_n$  for the hyperparameter  $k$  and plugs it in into the conditional posterior distribution given  $k$ , i.e.

$$\pi_{|\hat{k}_n}(\theta|\mathbf{Y}) = \pi_{|k}(\theta|\mathbf{Y}) \Big|_{k=\hat{k}_n},$$

which is the empirical Bayes posterior distribution.

Models in the form (3) are widely used in the Bayesian literature and under nonrestrictive assumptions the posterior distribution can optimally

recover the true parameter  $\theta_0$ . In more details, it is common to assume that the true parameter belongs to some regularity class  $\theta_0 \in \Theta^\beta$  with some unknown regularity hyper-parameter  $\beta$ . Then it was shown for instance in [1] that the hierarchical Bayes approach described above achieves optimal minimax contraction rate around the truth without using any additional information about its unknown regularity, leading to an adaptive procedure, in the frequentist sense. In this article our focus is on the quality of Bayesian uncertainty quantification done via credible balls from a frequentist perspective. There are two main properties of interest in a confidence set from a frequentist perspective: the frequentist coverage and the expectation of its size under  $P_{\theta_0}^{(n)}$  when  $\theta_0$  is assumed to be the true value of the parameter. In the literature the frequentist coverage properties of Bayesian credible sets constructed from sieve posteriors were only investigated for specific choice of priors and likelihoods, see for instance [2, 39, 49]. In this article we present a general approach under which we can simultaneously investigate the frequentist properties of the credible sets resulting from different choices of sieve priors and likelihoods.

We introduce some additional notations.

Let  $d(\cdot, \cdot)$  be a metric and  $B_k(\theta_1, u)$  the  $d(\cdot, \cdot)$  - ball in  $\Theta(k)$  with center  $\theta_1$  and radius  $u$ . Let  $\text{diam}(C, d(\cdot, \cdot))$  denote the  $d$ -diameter of the set  $C$ , i.e.

$$\text{diam}(C, d(\cdot, \cdot)) = \sup_{\theta_1, \theta_2 \in C} d(\theta_1, \theta_2).$$

We define

$$b(k) = \inf\{d^2(\theta_0, \theta), \theta \in \Theta(k)\}.$$

For simplicity we also extend the definition of the function  $b$  on  $[0, +\infty)$  by  $b(x) = b(k)$  for all  $x \in [k, k+1)$  and  $b(0) = +\infty$ . Note that we allow  $d(\cdot, \cdot)$  to depend on  $n$ , so that in this case  $b(k)$  also depends on  $n$ . This will be the case in particular in regression with fixed design and in the classification example. We also denote the Kullback-Leibler divergence

$$KL(\theta_0, \theta) = \frac{1}{n} E_{\theta_0}^{(n)} \left( \log \left( \frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}} \right) \right)$$

and the normalized variance of the log-likelihood with respect to  $E_{\theta_0}^{(n)}$  by

$$V(\theta_0, \theta) = \frac{1}{n} V_{\theta_0}^{(n)} \left( \log \left( \frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}} \right) \right).$$

**2. Main results.** In this Section we investigate the frequentist properties of Bayesian credible sets resulting from the hierarchical and the empirical Bayes procedures. We consider the general setting described in Section 1.1 and introduce nonrestrictive abstract conditions in the spirit of [18] under which credible sets have honest frequentist coverage and rate adaptive size. The derived results will be applied in Section 3 for various specific choices of sampling models and prior distributions.

Using the posterior distribution  $\pi(\theta|\mathbf{Y})$ , be it hierarchical or empirical, we construct the Bayes credible sets as balls centered around some estimator  $\hat{\theta}$  (typically the posterior mean, mode or median)  $\hat{C}(\alpha) = \{d(\theta, \hat{\theta}_n) \leq r_\alpha\}$  where  $\alpha \in (0, 1)$  and  $r_\alpha$  is the radius of the ball and satisfies

$$(5) \quad r_\alpha = \inf\{r, \pi(d(\theta, \hat{\theta}_n) \leq r | \mathbf{Y}) \geq 1 - \alpha\}.$$

In our analysis we also introduce some additional flexibility to the credible sets by allowing them to be blown up by a factor  $L > 0$  resulting in

$$\hat{C}(L, \alpha) = \{\theta : d(\theta, \hat{\theta}_n) \leq Lr_\alpha\}.$$

We show that these sets (for sufficiently large blow up factor  $L$ ) will have frequentist coverage tending to one and at the same time their size almost optimal in a minimax sense.

To control the frequentist coverage of  $\hat{C}(L, \alpha)$ , we need to restrict ourselves to a subset of  $\Theta$ , in a manner similar to [44], generalizing their idea outside the white noise model with empirical Gaussian prior process. We call general polished tail the condition required on the subclass of functions for which frequentist coverage can be obtained.

**DEFINITION 1.** *Let  $\theta \in \Theta$ , we say that  $\theta$  (or equivalently its associated bias function  $b(\cdot)$ ) satisfies the general polished tail condition associated to the metric  $d(\cdot, \cdot)$  if there exists  $k_0, R_0 > 1, \tau < 1$  such that*

$$(6) \quad b(kR) \leq \tau b(k), \quad \forall k \geq k_0.$$

*For given  $k_0, R_0$  and  $\tau$ , we denote by  $\Theta_0(R_0, k_0, \tau)$  the class of  $\theta \in \Theta$  satisfying (6).*

Note that in the case of the Gaussian white noise with  $\ell_2$ -norm  $b(k) = \sum_{j=k+1}^{\infty} \theta_j^2$ . The polished tail condition in [44] reads as

$$\sum_{i=N}^{\infty} \theta_i^2 \leq L \sum_{i=N}^{\rho N} \theta_i^2, \quad \forall N \geq N_0,$$

for some  $N_0, L, \rho > 0$  which is equivalent with our definition with  $k_0 = N_0$ ,  $\tau = (L - 1)/L$  and  $R = \rho$ . Our new definition, however, extends also to the case where the metric  $d(\cdot, \cdot)$  is substantially different from the  $\ell_2$  norm of  $\theta$ .

In the Gaussian white noise model with Gaussian prior, [44] shows that a key idea to obtain good coverage is that a trade-off between bias and variance is realized, so that the *correct* value of  $k$  (or set of values) is selected either under the posterior  $\pi_k(k|\mathbf{Y})$  or the empirical posterior distribution.

To generalize this idea in non Gaussian setups, let us define for each  $\theta_0 \in \Theta$ ,

$$(7) \quad \varepsilon_n^2(k) = b(k) + \frac{k \log n}{n}, \quad \text{and} \quad k_n = \inf\{k, b(k) \leq k \log n/n\},$$

and  $\mathcal{K}_n(M) = \{k; \varepsilon_n(k) \leq M\varepsilon_n(k_n)\}$ . Note that in these notations  $\theta_0$  is implicit since these quantities depend on  $\theta_0$ .

The generalization of the usual bias and variance trade-off is by obtaining a trade-off between the bias  $nb(k)$  and a prior penalization term induced by the prior mass of small neighbourhoods:  $\pi_{|k}(d(\theta_{[k]}^o, \theta) \leq u_n)$  where  $u_n = o(1)$  and  $\theta_{[k]}^o$  can be viewed as the projection of  $\theta_0$  on  $\Theta(k)$ , typically with respect of the KL-divergence. Then typically if  $u_n \asymp n^{-H}$  for some  $H > 0$ , then  $\log \pi_{|k}(d(\theta_{[k]}^o, \theta) \leq u_n) \asymp -k \log n$ , so that the set  $\mathcal{K}_n(M)$  corresponds to values of  $k$  for which this trade-off is achieved.

REMARK 1. *Note that if  $k \in \mathcal{K}_n(M)$ , then*

$$b(k) + \frac{k \log n}{n} \leq M^2 \left( b(k_n) + \frac{k_n \log n}{n} \right) \leq 2M^2 \frac{k_n \log n}{n},$$

*so that  $k \leq 2M^2 k_n$ . Moreover if  $b(\cdot) \in \Theta_0(R_0, k_0, \tau)$  and  $k_n > R_0$  then take  $m$  to be the smallest integer so that  $\tau^{-m} > 2R_0 M^2$  and consider any  $k \in \{k_n/R_0, \dots, k_n - 1\}$ , then*

$$b(k) \geq \frac{k \log n}{n} \geq \frac{k_n \log n}{R_0 n} \geq \frac{\varepsilon_n^2(k_n)}{2R_0},$$

$$b(k) = b(R_0^m R_0^{-m} k) \leq \tau^m b(R_0^{-m} k) \text{ and}$$

$$\varepsilon_n^2(\lfloor R_0^{-m} k \rfloor) \geq b(R_0^{-m} k) > 2R_0 M^2 b(k) > M^2 \varepsilon_n^2(k_n).$$

*Therefore we can conclude that  $\mathcal{K}_n(M) \subset \{R_0^{-m-1} k_n, \dots, 2M^2 k_n\}$ , since for every  $k \leq R_0^{-m-1} k_n$  there exists a  $k^* \in \{R_0^{-m-1} k_n, \dots, R_0^{-m}(k_n - 1)\}$  and  $\nu \in \mathbb{N}$  such that  $k = R_0^{-\nu} k^*$  and hence  $\varepsilon_n(k)^2 \geq b(k) \geq b(k^*) > M^2 \varepsilon_n^2(k_n)$ .*

We will show in Section 2.1 that in the hierarchical Bayes approach the posterior distribution concentrates on  $\mathcal{K}_n(M)$  for  $M$  large enough if the true parameter satisfies the general polished tail condition (6). A similar phenomenon occurs in the empirical Bayes for which we show that  $\hat{k}_n$  also belongs to  $\mathcal{K}_n(M)$  with high probability, see Section 2.2.

In both hierarchical and empirical Bayes setups we consider the following conditions on the prior  $\pi_{|k}$ .

Let  $\theta_{[k]}^o \in \Theta(k)$  be some point in  $\Theta(k)$ , typically  $\theta_{[k]}^o$  can be thought as the Kullback-Leibler projection of  $\theta_0$  onto  $\Theta(k)$ .

**C1** Prior on  $\theta|k$

$$\pi_{|k}(\theta) = \rho(k)g_k(\theta), \quad \rho(k) > 0$$

with

$$(8) \quad \sup_{\theta \in \Theta(k)} g_k(\theta) \leq C_\infty^k, \quad \inf_{B_k(\theta_{[k]}^o, \sqrt{k/n})} g_k(\theta) \geq C_B^k,$$

for some  $C_B, C_\infty > 0$ .

In the hierarchical prior case we also consider the following condition on the prior on  $k$ :

**C2** Prior on  $k$

$$(9) \quad e^{-c_1 t(k)k} \lesssim \pi_k(k) \lesssim e^{-c_2 t(k)k},$$

for  $t(k) = 1$  or  $\log k$  and for some  $c_1, c_2 > 0$ .

Define the following sets:

$$S_n(k, \kappa, \kappa') = \left\{ E_{\theta_0}^{(n)} \log \frac{p_{\theta_{[k]}^o}^{(n)}}{p_{\theta}^{(n)}} \leq \kappa k, E_{\theta_0}^{(n)} \left( \log \frac{p_{\theta_{[k]}^o}^{(n)}}{f_{\theta}} - E_{\theta_0}^{(n)} \log \frac{p_{\theta_{[k]}^o}^{(n)}}{p_{\theta}^{(n)}} \right)^2 \leq \kappa' k \right\},$$

$$\tilde{S}_n(k, \kappa, \kappa') = \left\{ KL(\theta_0, \theta) \leq \kappa \varepsilon_n(k)^2, V(\theta_0, \theta) \leq \kappa' \varepsilon_n(k)^2 \right\}.$$

In order to bound from below the frequentist coverage of  $\hat{C}(L, \alpha)$  and its size, we restrict ourselves to a subset of parameters  $\Theta_0 \subseteq \Theta(R_0, k_0, \tau)$  for some  $R_0, k_0, \tau$  for which we consider the following assumptions.

**A1** There exist  $\kappa, \kappa', \kappa_1 > 0$  such that for all  $\theta_0 \in \Theta_0$

$$B_k(\theta_0, \varepsilon_n(k_n)) \subset \tilde{S}_n(k_n, \kappa, \kappa'), \quad \pi_{|k_n}(B_{k_n}(\theta_0, \varepsilon_n(k_n))) \geq e^{-\kappa_1 k_n \log n}.$$

**A2(k)**

$$B_k(\theta_{[k]}^o, \sqrt{k/n}) \subset S_n(k, \kappa, \kappa'), \quad \pi_{|k}(B_k(\theta_{[k]}^o, \sqrt{k/n})) \geq e^{-\kappa_1 k \log n}.$$



**A3(k)** For all  $\varepsilon > 0$ , there exist sets  $\Theta_n(k)$  satisfying

$$\pi_{|k}(\Theta_n(k)^c) \leq e^{-(\kappa + \kappa_1 + 2/\varepsilon)n\varepsilon_n^2(k_n) - c_2 t(k_n)k_n}.$$

**A4** For all  $M$ , there exist constants  $M_0, B > 0$  such that

$$\sup_{\theta_0 \in \Theta_0} P_{\theta_0}^{(n)} \left( \max_{k \in \mathcal{K}_n(M)} \sup_{\Theta_n(k) \cap B_k(\theta_{[k]}^o, (M_0+1)\rho_n \varepsilon_n(k_n))} (\ell_n(\theta) - \ell_n(\theta_{[k]}^o) - Bk) > 0 \right) = o(1).$$

**A5(k)** There exist  $\zeta, c_0 > 0$ , such that for all  $\theta \in \Theta_n(k)$ , there exist test functions  $\varphi_n(\theta) \in [0, 1]$  satisfying

$$E_{\theta_0}^{(n)}(\varphi_n(\theta)) \leq e^{-c_0 n d^2(\theta_0, \theta)}, \quad \sup_{d(\theta', \theta) \leq \zeta d(\theta_0, \theta)} E_{\theta'}^{(n)}(1 - \varphi_n(\theta)) \leq e^{-c_0 n d^2(\theta_0, \theta)}$$

and there exists  $u_0 > 0$  such that for all  $u \geq (\sqrt{b(k)} \vee u_0 \sqrt{k \log n/n})$

$$(10) \quad N(\zeta u, \Theta_n(k) \cap \{u \leq d(\theta_0, \theta) \leq 2u\}, d(\cdot, \cdot)) \leq c_0 n u^2 / 2.$$

**A6(k)** For some  $J_0 \geq u_0 \vee 1$

$$\text{Vol}(\Theta_n(k) \cap B_k(\theta_0, J_0 \varepsilon_n(k))) \leq e^{-(\kappa + \kappa_1 + 3)n\varepsilon_n^2(k_n) - c_2 t(k_n)k_n}.$$

**A7(k)** Let  $\delta > 0$  be a small number. Then for all  $\tilde{\theta} \in \{\theta \in \Theta_n(k) : d(\theta, \theta_{[k]}^o) \leq M_0 \varepsilon_n(k_n)\}$ ,

$$\log \text{Vol}(B_k(\tilde{\theta}, \delta \sqrt{k/n})) - \log \text{Vol}(B_k(\theta_{[k]}^o, \sqrt{k/n})) \leq k \log(C\delta),$$

for some  $C > 0$ .

A brief explanation of the above conditions is in order. Assumptions **A3(k)**, **A5(k)** and the second part of **A1** are the standard remaining mass, entropy and prior small ball probability conditions, routinely used in the literature for determining the contraction rates of the posteriors. The first part of condition **A1** requires that locally the Kullback-Leibler divergence (and the variance of the log-likelihood ratio) is bounded from above by a multiple of the testing distance  $d(\cdot, \cdot)$ . Condition **A2(k)** is similar in spirit to **A1** but is slightly more involved. It requires that locally around  $\theta_{[k]}^o$ , which can be viewed as the Kullback-Leibler projection of  $\theta_0$  onto  $\Theta(k)$ , the testing distance  $d(\cdot, \cdot)^2$  bounds from above the difference between  $KL(\theta_0, \theta)$  and  $KL(\theta_0, \theta_{[k]}^o)$ , up to a multiplicative constant. The extra difficulty here lies in obtaining a sharp upper bound on  $KL(\theta_0, \theta) - KL(\theta_0, \theta_{[k]}^o)$  and not only on  $KL(\theta_0, \theta)$ . Both conditions **A1** and **A2(k)** are used to provide

lower bounds on the marginal density of the observations. In assumption **A4** we require that in each model  $\Theta(k)$ , the log likelihood ratio is uniformly bounded from above in small neighbourhoods of  $\theta_{[k]}^o$ . Condition **A6(k)** and condition **A7(k)** loosely speaking require an upper bound on  $d(\cdot, \cdot)$  - neighbourhoods. Condition **A6(k)** is quite common when an upper bound on the marginal density of model  $\Theta(k)$  is required, see for instance [31] or [36]. Condition **A7(k)** compares the volume of the small balls centred around the projection  $\theta_{[k]}^o$  and the centering point of the credible set (e.g. posterior mean). However, since the posterior mean is not known in advance the assumption considers balls centered around parameters in the neighbourhood of  $\theta_{[k]}^o$ , which contains the estimator with high probability. We require that in case the ball around  $\tilde{\theta}$  has substantially smaller radius than the ball centered around  $\theta_{[k]}^o$  then the volume of the ball is also substantially smaller. This is verified in particular when the distance  $d(\cdot, \cdot)$  behaves locally like the euclidian distance.

There are variants of the above conditions which can be considered following the usual variants which can be found in the literature on posterior concentration rates. Here we consider another version as well, which will be applied in the density estimation example with exponential families of priors and involves slicing the sets  $\Theta_n(k)$ .

**A5(k)'** The entropy condition (10) is replaced by: There exist a (possibly infinite) cover  $B_{n,j}(k)$  of the set  $\Theta_n(k) \cap \{\theta : d(\theta, \theta_0) \geq J_0 \varepsilon_n(k)\}$  such that

$$(11) \quad B_{n,j}(k) \subset \Theta_n(k) \cap \{d(\theta, \theta_0) > c(k, j) \varepsilon_n(k)\}$$

with

$$(12) \quad \sum_j \exp\left(-\frac{c_1}{2} n c(k, j)^2 \varepsilon_n(k)^2\right) \lesssim e^{-(\kappa + \kappa_1 + 1) n \varepsilon_n(k)^2},$$

where  $\kappa, \kappa_1$  are defined in assumption **A1** and

$$(13) \quad \log N(\zeta c(k, j) \varepsilon_n(k), B_{n,j}(k), d(\cdot, \cdot)) \leq \frac{c_1 c(k, j)^2 n \varepsilon_n(k)^2}{2}.$$

In the next subsections we show that under the above assumptions together with the general polished tail restriction the credible sets resulting both from the hierarchical and empirical Bayes procedures have optimal size and good frequentist coverage.

2.1. *Hierarchical Bayesian credible sets.* In this Section we present the results for the hierarchical prior defined by (4) satisfying assumptions **C1** and **C2**. We show that under the general polished tail condition and the assumptions introduced in the preceding section the credible set  $\hat{C}(L_n, \alpha)$  with  $L_n \gtrsim \sqrt{\log n}$  has good frequentist properties, i.e. it has good frequentist coverage and rate adaptive size on  $\Theta_0 = \mathcal{L}(R, k_0, \tau)$ ,  $R > 1$ ,  $k_0 \geq 1$  and  $\tau < 1$ .

**THEOREM 1.** *Under the prior satisfying **C1** and **C2**. Assume that **A1** holds and that there exist  $\theta_{[k]}^o \in \Theta(k)$  for all  $k \in \mathcal{K}_n(M)$  such that **A4** is verified. Let  $M$  and  $A$  satisfy:*

$$(14) \quad M^2 > 8(\kappa + \kappa_1 + 3 + c_2 \mathbf{1}_{t(k_n) = \log k_n})/c_0, \quad A \geq \frac{1}{c_1} \left( \frac{\kappa + \kappa_1 + 1}{t(k_n)} + \frac{c_2}{\log n} \right)$$

*Assume that for all  $\{k \notin \mathcal{K}_n(M) : k \leq Ak_n \log n\}$ , **A3(k)**, **A5(k)** hold, that **A6(k)** holds for all  $\{k \notin \mathcal{K}_n(M) : M^2 k_n/2 \leq k \leq Ak_n \log n\}$  and that for all  $k \in \mathcal{K}_n(M)$  **A2(k)** and **A7(k)** hold. If the centering point satisfies that for all  $\varepsilon > 0$  there exists  $M_\varepsilon > 0$  such that*

$$(15) \quad \sup_{\theta_0 \in \Theta_0} P_{\theta_0}^{(n)} \left( d(\theta_0, \hat{\theta}_n) > M_\varepsilon \varepsilon_n(k_n) \right) \leq \varepsilon,$$

*then there exists a constant  $L_{\varepsilon, \alpha} > 0$  such that*

$$(16) \quad \liminf_n \inf_{\theta_0 \in \Theta_0} P_{\theta_0}^{(n)} \left( \theta_0 \in \hat{C}(L_{\varepsilon, \alpha} \sqrt{\log n}, \alpha) \right) > 1 - 2\varepsilon.$$

The proof of the theorem is deferred to Section 4.1. A key step in the proof is understanding the asymptotic distribution of  $\pi_k(k|\mathbf{Y})$ . In particular we show that the posterior distribution accumulates most of its mass on  $\mathcal{K}_n(M)$ , where the correct trade-off between bias and prior-penalization or complexity (equivalent to the variance term in the Gaussian setup) is achieved. This is presented in the following lemma:

**LEMMA 1.** *Consider a prior satisfying **C1** and **C2**. Assume that **A1** and **A4** are verified, that for all  $k \notin \mathcal{K}_n(M)$ ,  $k \leq Ak_n \log n$ , **A3(k)**, **A5(k)** hold and that for all  $M^2 k_n/2 \leq k < Ak_n \log n$  **A6(k)** holds. Furthermore, let  $M$  satisfy (14). Then*

$$\sup_{\theta_0 \in \Theta_0} E_{\theta_0}^{(n)} (\pi_k(k \notin \mathcal{K}_n(M)|\mathbf{Y})) = o(1).$$

The proof is presented in Section 4.2.

In the following lemma we show that  $\varepsilon_n(k_n)$  corresponds to the posterior concentration rates, hence  $\hat{\theta}_n$  can be any random point of the posterior distribution or depending on  $d(.,.)$  the posterior mean or other posterior summary.

LEMMA 2. *Assume that for all  $k \leq Ak_n \log n$ , **A3(k)**, **A5(k)** hold, that for all  $M^2 k_n/2 \leq k < Ak_n \log n$  **A6(k)** holds and that for all  $k \in \mathcal{K}_n(M)$  **A2(k)** holds. Then there exists  $M_1 > 0$  such that*

$$\sup_{\theta_0 \in \Theta_0} E_{\theta_0}^{(n)}(\pi(d(\theta, \theta_0) \geq M_1 \varepsilon_n(k_n) | \mathbf{Y})) = o(1).$$

The proof of Lemma 2 is presented in Section 4.3.

Finally we show that the radius of the credible set is bounded from above by  $\varepsilon_n(k_n)$ .

COROLLARY 1. *Under the assumptions of Lemma 2 and (15) we have for all  $\varepsilon \in (0, 1/2)$  that*

$$\inf_{\theta_0 \in \Theta_0} P_{\theta_0}^{(n)}(\text{diam}(\hat{C}(1, \alpha), d(.,.)) \leq (M_\varepsilon + M_1) \varepsilon_n(k_n)) \geq 1 - 2\varepsilon.$$

The lemma is a straightforward consequence of assumption (15) and Lemma 2.

**2.2. Empirical Bayes approach.** An alternative approach to endow the hyper-parameter by a prior is to estimate it from the data directly and plug in this estimator into the posterior distribution. One of the most commonly used approach is the maximum marginal likelihood empirical Bayes approach, where one estimates the hyperparameter with the maximizer of the marginal likelihood function

$$(17) \quad \hat{k}_n = \arg \max_k \int_{\Theta(k)} e^{\ell_n(\theta)} \pi_{|k}(\theta) d\theta,$$

where  $\ell_n(\theta)$  denotes the loglikelihood function. This empirical Bayes technique is closely related to the hierarchical Bayes approach, however, in certain situations they can have substantially different behaviour, see for instance [29, 36].

In the empirical Bayes approach we construct the credible set similarly to the hierarchical Bayes case, i.e. we consider a  $d$ -ball around the empirical Bayes estimator  $\hat{\theta}_n$  (typically posterior mean or mode)

$$(18) \quad \hat{C}_{\hat{k}_n}(L, \alpha) = \{\theta : d(\theta, \hat{\theta}_n) \leq L r_\alpha(\hat{k}_n)\},$$

where  $L > 0$  is a blow up factor and

$$(19) \quad \pi_{|\hat{k}_n}(d(\theta, \hat{\theta}_n) \leq r_\alpha(\hat{k}_n) | \mathbf{Y}) = 1 - \alpha,$$

where  $\alpha \in (0, 1)$  is typically small. We show that these sets have similar size as the hierarchical Bayes credible sets and good frequentist coverage under the general polished tail condition (6).

**THEOREM 2.** *Under the assumption of Theorem 1 and  $|\mathcal{K}| \lesssim n^a$  for some  $a > 0$ , we have for every  $\varepsilon, \alpha \in (0, 1)$  and  $\hat{\theta}_n$  satisfying*

$$\sup_{\theta_0 \in \Theta_0} P_{\theta_0}^{(n)} \left( d(\theta_0, \hat{\theta}_n) > M_\varepsilon \varepsilon_n(k_n) \right) \leq \varepsilon$$

that there exists a constant  $L_{\varepsilon, \alpha} > 0$  such that  $\hat{C}_{\hat{k}_n}(L_{\varepsilon, \alpha} \sqrt{\log n}, \alpha)$  verifies (16).

Furthermore, there exists  $K_\varepsilon > 0$  such that

$$\inf_{\theta_0 \in \Theta_0} P_{\theta_0}^{(n)} \left( \text{diam}(\hat{C}_{\hat{k}_n}(1, \alpha), d(\cdot, \cdot)) \leq (M_\varepsilon + M_1) \varepsilon_n(k_n) \right) \geq 1 - 2\varepsilon.$$

The proof is deferred to Section 4.4.

### 3. Application to various models.

**3.1. Application to fixed or random design regression.** In this section we consider the fixed design regression model and investigate the behaviour of Bayesian credible sets based on sieve priors. Assume that we observe the sequence  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$  satisfying

$$(20) \quad Y_i = f_0(x_i) + \sigma Z_i, \quad x_i \in [0, 1], \quad i = 1, 2, \dots, n,$$

where  $Z_i$  are iid standard normal random variables,  $\sigma = 1$  for simplicity and  $x_1, x_2, \dots, x_n$  are fixed (or random) design points.

Next we consider the basis  $\phi_1(x), \phi_2(x) \dots$  in  $L_2[0, 1]$  and assume that  $\phi_i = (\phi_i(x_1), \phi_i(x_2), \dots, \phi_i(x_n))^T \in \mathbb{R}^n$ ,  $i = 1, 2, \dots, n$  forms a basis in  $\mathbb{R}^n$ . Note that every  $f \in L_2[0, 1]$  can be written in the form  $f(x) = f_\theta(x) = \sum_{i=1}^{\infty} \theta_i \phi_i(x)$  and we assume that the true function  $f_0$  belongs to a Sobolev smoothness class  $S^\beta(M)$ , i.e.

$$f_{\theta_0} \in S^\beta(M) = \{f_\theta : \sum_{i=1}^{\infty} \theta_i i^{2\beta} \leq M\},$$

for some  $\beta, M > 0$ .

Next, for any  $k \leq n$  we introduce the notation  $\Phi_k = (\phi_1, \phi_2, \dots, \phi_k) \in \mathbb{R}^{n \times k}$ . Let  $d_n(f_1, f_2)^2 = \frac{1}{n} \sum_{i=1}^n (f_1(x_i) - f_2(x_i))^2$  be the empirical  $L_2$ -norm between the vectors  $f_{1,n}$  and  $f_{2,n}$  where  $f_{j,n} = (f_j(x_1), \dots, f_j(x_n))$ . Denote by  $\theta_{[k]}^o$  the empirical  $L_2$ -norm projection of  $f_{0,n} = (f_0(x_1), \dots, f_0(x_n))^T$  to the space  $\{\Phi_k \theta : \theta \in \mathbb{R}^k\}$ . Then defining  $b(k)$  in terms of the (pseudo metric)  $d_n(\cdot, \cdot)$  leads to  $b(k) = d_n(f_0, \Phi_k \theta_{[k]}^o)^2$  the approximation error of the true function with a  $k$  dimensional projection. Assume furthermore that there exists a constant  $C_0 > 0$  and a sequence  $K_n$  going to infinity such that

$$(21) \quad C_0^{-1} I_k \leq \frac{\Phi_k^T \Phi_k}{n} \leq C_0 I_k, \quad \forall k \leq K_n.$$

REMARK 2. *The above assumptions on the choice of the basis functions  $\phi_j(x) \in L_2[0, 1]$  and the design points  $x_1, x_2, \dots, x_n$  are very mild and standard. There are many suitable choice of basis satisfying these properties. Orthonormal bases in  $\mathbb{R}^n$ , such as the discrete wavelet bases relative to the design points satisfy (21) with  $K_n = n$ , some orthonormal bases in  $L_2$  will satisfy (21) for some finite value  $K_n$ . In the case of the Fourier basis for instance, (21) is valid as soon as  $K_n = o(n)$ .*

REMARK 3. *Note that in the case of random design, with known distribution  $\nu$ , under boundedness condition on the  $\phi_j$ 's which form an orthonormal system of  $L_2(\nu)$ ,*

$$(22) \quad |E_\nu(\phi_j \phi_l) - \mathbf{1}_{j=l}| \leq M \sqrt{\log K_n} / \sqrt{n}$$

*with probability going to 1, uniformly over  $l, j \leq K_n$ . Hence if  $K_n \sqrt{\log K_n} / \sqrt{n} = o(1)$ , (21) is verified  $\nu$ -almost surely.*

Denote by  $\Theta_{0,n} = \Theta_0 \cap \{\theta_0; b(K_n) \leq \delta K_n \log n/n\}$  for some  $\delta < 1 \wedge C_0$  and consider  $\theta_0 \in \Theta_{0,n}$ . To understand better the meaning of the restriction  $\theta_0 \in \Theta_{0,n}$ , assume that  $\sum_{j=1}^\infty |\theta_{0,j}| < +\infty$ . If (21) is true for all  $1 \leq k \leq n$ , then writing  $\Delta_k = f_0 - \sum_{j=1}^k \theta_{0,j} \phi_j$  we have that  $\|\Delta_k\|_\infty = o(1)$  as  $k$  goes to infinity, which given that  $b(k) \leq \|\Delta_k\|_\infty^2$  implies that there exists  $K_n \geq 1$  such that  $b(K_n) \leq \delta K_n \log n/n$  for all  $n \geq 2$  and  $\delta > 0$ . Hence  $\{\theta; \|\theta\|_1 < +\infty\} \cap \Theta_0 \subset \Theta_{0,n}$ . However if (21) is only true for some finite sequence  $K_n$  going to infinity, then  $\Theta_{0,n}$  will typically be more constraint. For instance, assume a Sobolev  $\beta$  regularity on  $\theta_0$ , we can bound, if  $\beta > 1/2$ ,  $b(K_n) \leq \|\Delta_{K_n}\|_\infty^2 \lesssim K_n^{-2(\beta-1/2)}$  so that  $b(K_n) \leq \delta K_n \log n/n$  if  $K_n \gtrsim (n/\log n)^{1/(2\beta)}$ . For instance if  $K_n \asymp n/\log n$ ,  $\beta > 1/2$  is enough, while

if  $K_n \asymp \sqrt{n/\log n}$  then one needs  $\beta > 1$ . The upper bound  $K_n^{-2(\beta-1/2)}$  is independent of the design and the chosen basis and can be improved in particular cases.

In the random design case with distribution  $\nu$  and with bounded orthonormal basis:  $\max_j \|\phi_j\|_\infty < +\infty$ , one has

$$\begin{aligned} \nu(d_n^2(\Delta_{K_n}) > C\|\Delta_{K_n}\|_2^2) &= \nu\left(\sum_{i=1}^n \left(\sum_{j=K_n+1}^{\infty} \theta_{0,j} \phi_j(x_i)\right)^2 > nC\|\Delta_{K_n}\|_2^2\right) \\ &\leq \frac{E_\nu\left((\sum_{j=K_n+1}^{\infty} \theta_{0,j} \phi_j(X))^2\right)}{C\|\Delta_{K_n}\|_2^2} \leq \frac{1}{C} \end{aligned}$$

and  $b(K_n) \leq d_n^2(\Delta_{K_n}) \leq C\|\Delta_{K_n}\|_2^2 \lesssim K_n^{-2\beta}$  with large probability.

Then we define the prior distribution on the regression function  $f$  with hyper-parameter  $k$  by endowing the sequence of coefficients  $\theta$  with the standard sieve prior, i.e.

$$\theta = (\theta_1, \dots, \theta_k) | k \sim \prod_{i=1}^k g(\theta_i),$$

where  $g(\cdot)$  satisfies the standard assumption

$$(23) \quad \int_{\mathbb{R}} e^{s_0|x|^p} g(x) dx = a < \infty$$

and by either endowing  $k$  with a hyper-prior  $\pi_k(k)$  satisfying (9) or estimating it by the MMLE (17). These type of priors were considered for instance in [1] and [36], where it was shown that the corresponding hierarchical and empirical Bayes posterior distributions achieve adaptive contraction rate around the true function  $f_0$ . The frequentist behaviour of the Bayesian credible sets in context of the regression model was investigated only in a few papers [39, 40, 49] for specific choices of the prior and using direct computations. Here we consider both the hierarchical Bayes credible set

$$\hat{C}(L\sqrt{\log n}, \alpha) = \{f_\theta : d_n(f_\theta, f_{\hat{\theta}}) \leq L\sqrt{\log n r_\alpha}\},$$

with  $\pi(f_\theta : d_n(f_\theta, f_{\hat{\theta}}) \leq r_\alpha | \mathbf{Y}) = 1 - \alpha$  and  $f_{\hat{\theta}}$  satisfying assumption (15) and the MMLE empirical Bayes credible set defined along the same lines. By applying Theorems 1 and 2 together with Corollary 1 we can verify that both credible sets have good frequentist coverage and (almost) rate adaptive size under the general polished tail assumption.

PROPOSITION 1. *Consider the fixed design regression model (20) with  $f_0 \in S^\beta(L_0)$  for some  $\beta \geq \beta_0 > 1/2$  and assume that condition (21) is satisfied with  $K_n > n^{\frac{\beta_0}{(1+2\beta_0)(\beta_0-1/2)}}$ . Denote both the hierarchical Bayes and empirical Bayes credible sets, centered around any estimator  $f_{\hat{\theta}_n}$  satisfying (15) by  $\hat{C}_n(L\sqrt{\log n}, \alpha)$ . Then  $\hat{C}_n(L\sqrt{\log n}, \alpha)$  has (up to a  $\log n$  factor) rate adaptive size and frequentist coverage tending to one under the general polished tail assumption (6), i.e. for every  $\varepsilon > 0$  there exist a large enough  $L, C > 0$  such that*

$$\liminf_n \inf_{f_{\theta_0}: \theta_0 \in \Theta_{0,n} \cap S^{\beta_0}} P_{f_{\theta_0}}^{(n)}(f_0 \in \hat{C}_n(L\sqrt{\log n}, \alpha)) \geq 1 - \varepsilon,$$

$$\liminf_n \inf_{\beta > \beta_0} \inf_{f_{\theta_0} \in S^\beta(M)} P_{f_{\theta_0}}^{(n)}\left(\text{diam}(\hat{C}_n(\sqrt{\log n}, \alpha), d_n) \leq C\sqrt{\log n} \left(\frac{n}{\log n}\right)^{-\frac{\beta}{1+2\beta}}\right) \geq 1 - \varepsilon.$$

PROOF. The proof of the Proposition is given in Section A.1.  $\square$

REMARK 4. *Assumption (15) on the estimator is very mild, for instance a typical draw from the posterior distribution satisfies it, see the comment above Lemma 2. Furthermore, standard estimators, for instance the posterior mean also satisfies this assumption, see for instance [1].*

3.2. *Application to density estimation using histogram priors.* In this section we consider the density estimation model, i.e. we assume to observe  $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_n\}$  iid samples from a true density function  $p_0$  and our goal is to recover this density. We assume that  $p_0$  is continuous, bounded from below by  $c_0$  and from above by  $C_0$ . Furthermore we assume that it belongs to a Holder smoothness class  $\mathcal{H}^\beta(L_0)$  for some  $\beta \in (0, 1]$ .

We investigate the Bayesian approach using histogram prior distributions, see for instance [13, 36, 38]. In other words let  $\Theta(k)$  denote the collection of  $k$ -bins random histogram where the bins are regular :  $[(j-1)/k, j/k)$ ,  $j = 1, \dots, k$ ,

$$(24) \quad p_\theta(x) = k \sum_{j=1}^k \theta_j \mathbf{1}_{I_j}(x), \quad \theta_j \geq 0, \quad \sum_{j=1}^k \theta_j = 1.$$

We therefore identify  $\Theta(k)$  with the  $k$ -dimensional simplex  $\mathcal{S}_k = \{x \in [0, 1]^k; \sum_{i=1}^k x_i = 1\}$ . First we endow the hyper-parameter  $k$  with either a Poisson  $\mathcal{P}(\lambda)$  or a Geometric  $\mathcal{G}(q)$  hyper-prior with  $\lambda > 0$  and  $0 < q < 1$ . Given  $k$  consider a Dirichlet prior  $\mathcal{D}(1, \dots, 1)$  on  $(\theta_1, \dots, \theta_k)$ , i.e. the hier-



archical prior  $\pi$  on the densities takes the form

$$\begin{aligned}\theta &= (\theta_1, \dots, \theta_k) | k \sim \mathcal{D}(1, \dots, 1) \\ k &\sim \text{Geom}(q) \text{ or } \text{Pois}(\lambda),\end{aligned}$$

for some fixed  $q \in (0, 1)$  or  $\lambda > 0$ . Alternatively we apply the MMLE  $\hat{k}_n$  for the hyper-parameter  $k$  and then consider the Dirichlet prior  $\mathcal{D}(1, \dots, 1)$  on  $(\theta_1, \dots, \theta_{\hat{k}_n})$ .

Then we consider the hierarchical Bayes credible set

$$\hat{C}(L\sqrt{\log n}, \alpha) = \{p_\theta : h(p_\theta, p_{\hat{\theta}}) \leq L\sqrt{\log nr_\alpha}\},$$

with  $h(\cdot, \cdot)$  the hellinger distance,  $p_{\hat{\theta}_n}$  satisfying assumption (15) with  $d(\theta_1, \theta_2) = h(p_{\theta_1}, p_{\theta_2})$  and  $\pi(p_\theta : h(p_\theta, p_{\hat{\theta}_n}) \leq r_\alpha | \mathbf{Y}) = 1 - \alpha$ . The empirical Bayes credible set  $\hat{C}_{\hat{k}_n}(L\sqrt{\log n}, \alpha)$  is defined along the same lines. Applying again Theorems 1 and 2 together with Corollary 1 we can verify that both credible sets have good frequentist coverage and (almost) rate adaptive size under the general polished tail assumption.

**PROPOSITION 2.** *Consider the density estimation model with histogram priors (24) and assume that  $p_0 \in \mathcal{H}^\beta(M)$  for some  $\beta \in (0, 1]$  and it is bounded away from zero. Then both the hierarchical Bayes and empirical Bayes credible sets with centering point  $p_{\hat{\theta}}$  satisfying (15) have (up to a  $\log n$  factor) rate adaptive size and frequentist coverage tending to one under the polished tail assumption (6), i.e. for every  $\varepsilon > 0$  there exist  $L, C > 0$  such that*

$$\begin{aligned}\liminf_n \inf_{p_0 \in \Theta_0} P_{p_0}^{(n)}(p_0 \in \hat{C}_n(L\sqrt{\log n}, \alpha)) &\geq 1 - \varepsilon, \\ \liminf_n \inf_{\beta \in (0, 1]} \inf_{p_0 \in \mathcal{H}^\beta(L_0)} P_{p_0}^{(n)}\left(\text{diam}(\hat{C}_n(L\sqrt{\log n}, \alpha), h) \leq C\sqrt{\log n} \left(\frac{n}{\log n}\right)^{-\frac{\beta}{1+2\beta}}\right) &\geq 1 - \varepsilon,\end{aligned}$$

where  $\hat{C}_n(L\sqrt{\log n}, \alpha)$  could either denote the hierarchical or the empirical Bayes credible sets.

**PROOF.** The proposition is verified in Section A.2. □

**3.3. Application to density estimation with exponential families of prior.** In this subsection we consider again the density estimation problem on  $[0, 1]$ , i.e. we assume that we observe independent and identically distributed draws  $Y_1, Y_2, \dots, Y_n$ , denoted by  $\mathbf{Y}$ , from a distribution with density function  $f_0$

(with respect to the Lebesgue measure). Then we assume that the true density can be written as an infinite dimensional exponential distribution

(25)

$$f_0(x) = f_{\theta_0}(x) = \exp \left( \sum_{j=1}^{\infty} \theta_{0,j} \phi_j(x) - c(\theta_0) \right), \quad e^{c(\theta_0)} = \int_0^1 \exp \left( \sum_{j=1}^{\infty} \theta_{0,j} \phi_j(x) \right) dx$$

for some  $\theta_0 \in \ell_2$ . This model is also known as the log-linear model. Furthermore we also assume that  $\|\log f_0\|_{\infty} < +\infty$ , that  $\phi_j(x)$ ,  $j = 1, 2, \dots$  forms an orthonormal basis (together with  $\phi_0(x) \equiv 1$ ) and therefore satisfies  $\int_0^1 \phi_j(x) dx = 0$  for all  $j \geq 1$ , and that  $\theta_0 \in \mathcal{S}^{\beta}(L_0)$  for some  $\beta, L_0 > 0$ .

Then we define the prior distribution on the densities with hyper-parameter  $k$  by endowing the sequence  $\theta$  in the log-linear model with the standard sieve prior, i.e.

$$\theta = (\theta_1, \dots, \theta_k) | k \sim \prod_{i=1}^k g(\theta_i)$$

with  $g(\cdot)$  satisfying (23). Then we either endow the hyper-parameter  $k$  with a hyper-prior  $\pi_k(k)$  defined in (9) or by estimating it from the data by the MMLE.

These type of priors were considered for instance in [1, 33, 34, 36, 47, 48], where rate adaptive posterior contraction rates were shown. However, the reliability of Bayesian uncertainty quantification in this model was not investigated yet in the literature.

By using the corresponding posterior distribution we construct the hierarchical credible set as

$$\hat{C}(L\sqrt{\log n}, \alpha) = \{f_{\theta} : h(f_{\theta}, f_{\hat{\theta}_n}) \leq L\sqrt{\log n r_{\alpha}}\},$$

where  $h(\cdot, \cdot)$  denotes the hellinger distance,  $\pi(f_{\theta} : h(f_{\theta}, f_{\hat{\theta}_n}) \leq r_{\alpha} | \mathbf{Y}) = 1 - \alpha$  and  $f_{\hat{\theta}_n}$  an arbitrary estimator satisfying (15) with  $d(\cdot, \cdot) = h(\cdot, \cdot)$ . The construction of the empirical Bayes credible set  $\hat{C}_{\hat{k}_n}(L \log n, \alpha)$  goes similarly. Using again Theorems 1 and 2 together with Corollary 1 we can verify that the preceding credible sets have good frequentist coverage and (almost) rate adaptive size under the general polished tail assumption.

**PROPOSITON 3.** *Consider the log-linear model (25) then both the hierarchical and empirical Bayes credible sets have (up to a  $\log n$  factor) rate*

adaptive size and frequentist coverage tending to one under the general polished tail assumption (6), i.e. for every  $\beta_0 > 1/2$  and  $\varepsilon > 0$  there exist  $L, C > 0$  such that

$$\liminf_n \inf_{f_{\theta_0}: \theta_0 \in \Theta_0 \cap \mathcal{S}^{\beta_0}(L_0)} P_{f_{\theta_0}}^{(n)}(f_{\theta_0} \in \hat{C}_n(L\sqrt{\log n}, \alpha)) \geq 1 - \varepsilon,$$

$$\liminf_n \inf_{\beta > \beta_0} \inf_{f_{\theta_0} \in \mathcal{S}^\beta(L_0)} P_{f_{\theta_0}}^{(n)}\left(\text{diam}(\hat{C}_n(L\sqrt{\log n}, \alpha), h) \leq C\sqrt{\log n} \left(\frac{n}{\log n}\right)^{-\frac{\beta}{1+2\beta}}\right) \geq 1 - \varepsilon,$$

where  $\hat{C}_n(L\sqrt{\log n}, \alpha)$  denotes either the hierarchical or the empirical Bayes credible set.

PROOF. The proof of the Proposition is given in Section A.3. □

3.4. *Application to nonparametric classification.* In this section we apply our general theorem to the nonparametric classification (or also known as binary regression) model. We assume to observe the sequence  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n) \in \{0, 1\}^n$  satisfying

$$(26) \quad P(Y_i = 1|x_i) = q(x_i), \quad \text{for some } q : [0, 1] \mapsto (0, 1),$$

and write  $\mu(x) = e^x/(1 + e^x)$ , the logistic link function.

We assume that under the true distribution associated to  $q_0$ ,  $f_0 = \mu^{-1}(q_0) \in \mathcal{S}^\beta(L_0)$ , with unknown smoothness parameter  $\beta > 0$ . In the Bayesian approach one endows the nonparametric function  $f$  with a prior distribution resulting a prior on the binary regression function  $q$ . The theoretical properties of the Bayesian approach in the present model was investigated for instance in [18] with linear function  $f$ , in [47] with Gaussian process priors on the nonparametric function  $f$  and in [23] in context of classification of the nodes of large graphs. In the preceding papers optimal (and adaptive) posterior contraction rates were derived. However, the coverage properties of Bayesian credible sets remained unknown. We tackle this until now unanswered question by applying our general, abstract theorem.

In our analysis we consider again the popular sieve prior. For given  $k$  we introduce the parametrization

$$f_\theta(x_i) = \sum_{j=1}^k \theta_j \phi_j(x_i) = \Phi_k(x_i) \theta,$$

with  $\theta = (\theta_1, \dots, \theta_k)^T \in \mathbb{R}^k$  and  $\Phi_k(x_i) = (\phi_1(x_i), \phi_2(x_i), \dots, \phi_k(x_i))$ , as in Section 3.1. We note that in this case the models are nested  $\Theta(k+1) \subset \Theta(k)$

and hence  $b(k+1) \leq b(k)$ . Then we endow  $\theta_j \sim g(\cdot)$  with  $g(\cdot)$  satisfying (23) and  $k$  with the prior  $\pi_k(k)$  defined in (9) resulting the two level hierarchical prior  $\pi(\cdot)$

$$\theta = (\theta_1, \dots, \theta_k) | k \sim \prod_{i=1}^k g(\theta_i), \quad k \sim \pi_k(k).$$

Alternatively, we estimate  $k$  using the MMLE and then plug it in into the prior for  $\theta$  given  $k$ .

Then we consider credible balls in terms of  $q(x) = \mu(f(x))$ , and the average (empirical) Hellinger metric

$$h_n^2(q_1, q_2) = \frac{1}{n} \sum_{i=1}^n h_b^2(q_1(x_i), q_2(x_i)),$$

$$h_b(q_1(x_i), q_2(x_i)) = (\sqrt{q_1(x_i)} - \sqrt{q_2(x_i)})^2 + (\sqrt{1 - q_1(x_i)} - \sqrt{1 - q_2(x_i)})^2.$$

We center the ball at  $\hat{q}(x)$  the posterior mean of  $\mu(f_\theta(x))$ . By convexity and boundedness of  $h_n^2$

$$h_n^2(\hat{q}, q_0) \leq E^\pi (h_n^2(q_\theta, q_0)) \leq \varepsilon_n^2 + 2\pi (h_n^2(q_\theta, q_0) > \varepsilon_n^2 | \mathbf{Y}) \leq \varepsilon_n^2(1 + o(1)),$$

where  $E^\pi$  denotes the expectation with respect to the posterior, as soon as  $\varepsilon_n^2 \gg (n\varepsilon_n^2)^{-1}$ .

Moreover, note that  $h_n^2(q_1, q_2) \leq d_n^2(f_1, f_2)$  with  $f_j(x) = \mu^{-1}(q_j(x))$ ,  $j = 1, 2$ . Similarly to before, to understand the coverage properties of the credible balls, we need to study the bias function  $b(k)$  with respect to the (semi) metric  $h_n$ . Assume  $\theta_0 \in \mathcal{S}^\beta(L_0)$  for  $\beta \geq \beta_0 > 1/2$  and  $L_0 > 0$ . Denote by  $\tilde{b}(\cdot)$  the bias function associated to  $d_n(f_{\theta_0}, f_\theta)$  and studied in Section 3.1. Assume that  $K_n$  satisfies  $\tilde{b}(K_n) \leq \delta K_n \log n/n$  for some small enough  $\delta$ . Then since  $b(K_n) \leq \tilde{b}(K_n)$ ,  $b(K_n) \leq \delta K_n \log n/n$ . Denote  $f_{0,[k]} = \sum_{j=1}^k \theta_{0,j} \phi_j$ ,  $f_{0,[k],n} = (f_{0,[k]}(x_1), \dots, f_{0,[k]}(x_n))^T$ , and  $\Delta_{n,k} = f_{0,n} - f_{0,[k],n}$ , then for  $k \leq K_n$  we have  $\|f_0\|_\infty \vee \|f_{0,[k]}\|_\infty \leq \max_j \|\phi_j\|_\infty \|\theta_0\|_1 < +\infty$ . As in Section 3.1, we have that  $k_n \leq \tilde{k}_n := \min\{k, \tilde{b}(k) \leq k \log n/n\}$  with  $\tilde{k}_n \lesssim (n/\log n)^{1/(2\beta+1)}$ . The discussion on the feasibility of the constraint  $\tilde{b}(K_n) \leq \delta K_n \log n/n$  is similar to that of Section 3.1.

By using the corresponding posterior distributions we construct the hierarchical credible sets

$$\hat{C}(L\sqrt{\log n}, \alpha) = \{q_\theta(\cdot), h_n(q_\theta, \hat{q}) \leq L\sqrt{\log n} r_\alpha\}$$

with  $\pi(q_\theta(\cdot) : h_n(q_\theta, \hat{q}) \leq r_\alpha | \mathbf{Y}) = 1 - \alpha$  and similarly for  $\hat{C}_{\hat{k}}(L\sqrt{\log n}, \alpha)$ . By applying our main Theorems 1 and 2 and Corollary 1 we show that

under the polished tail assumption (6) both of the credible sets have good frequentist behaviour.

PROPOSITION 4. *Consider the classification model given in (26) with  $f_0 \in S^\beta(L_0)$ ,  $\beta \geq \beta_0 > 1/2$  and  $K_n \gg n^{\frac{1}{2(\beta_0-1/2)}}$ . Then the hierarchical and empirical Bayes credible sets  $\hat{C}_n(L\sqrt{\log n}, \alpha)$  - denoting either  $\hat{C}(L\sqrt{\log n}, \alpha)$  in the hierarchical approach or  $\hat{C}_{k_n}(L\sqrt{\log n}, \alpha)$  in the empirical approach - have (up to a  $\log n$  factor) rate adaptive size and frequentist coverage tending to one under the polished tail assumption, i.e. for every  $\varepsilon > 0$  there exist constants  $L, C > 0$  such that*

$$\liminf_n \inf_{q_0: \theta_0 \in \Theta_0 \cap S^{\beta_0}(L_0)} \inf_{\theta_0 \in S^\beta(L_0)} P_{q_0}^{(n)}(q_0 \in \hat{C}_n(L\sqrt{\log n}, \alpha)) \geq 1 - \varepsilon,$$

$$\liminf_n \inf_{\beta \geq \beta_0} \inf_{q_0: \theta_0 \in S^\beta(L_0)} P_{q_0}^{(n)}\left(\text{diam}(\hat{C}_n(L\sqrt{\log n}, \alpha), h) \leq C\sqrt{\log n} \left(\frac{n}{\log n}\right)^{-\frac{\beta}{1+2\beta}}\right) \geq 1 - \varepsilon.$$

PROOF. The proof of the proposition is deferred to Section A.4.  $\square$

REMARK 5. *The same coverage and contraction rate results can be shown for the empirical  $L_2$ -distance  $d_n(f_1, f_2)$  and for the  $L_2$  distance  $\|\theta_1 - \theta_2\|_2$  as well.*

REMARK 6. *Using similar computations other link functions  $\mu : \mathbb{R} \mapsto (0, 1)$  could be also considered satisfying*

$$(27) \quad \frac{\mu'(x)^2}{\mu(x)(1 - \mu(x))} \leq K_0,$$

for some  $K_0 > 0$ .

#### 4. Proof of the main results.

4.1. *Proof of Theorem 1.* Theorem 1 is a simple consequence of the following Lemma which allows to control the prior mass of neighbourhoods of  $\hat{\theta}_n$ .

LEMMA 3. *Consider the same assumptions as in Theorem 1. Let  $\rho_n = \delta/\sqrt{\log n}$  for some sufficiently small  $\delta > 0$  and  $\hat{\theta}_n$  satisfying (15). Then*

$$\sup_{\theta_0 \in \Theta_0} E_{\theta_0}^{(n)}\left(\pi(d(\theta, \hat{\theta}_n) \leq \rho_n \varepsilon_n(k_n) | \mathbf{Y})\right) = o(1).$$

The proof of Lemma 3 is presented in Section 4.1.1. We now give the proof of Theorem 1.

PROOF OF THEOREM 1. Let  $L_n = L_0\sqrt{\log n}$  and  $\varepsilon_n = \varepsilon_n(k_n)$ , then

$$\begin{aligned} P_{\theta_0}^{(n)}\left(\theta_0 \in \widehat{C}(L_n, \alpha)\right) &= P_{\theta_0}^{(n)}\left[d(\theta_0, \hat{\theta}_n) \leq L_n r_\alpha\right] \\ &\geq P_{\theta_0}^{(n)}\left[d(\theta_0, \hat{\theta}_n) \leq L_n r_\alpha \mid d(\theta_0, \hat{\theta}_n) \leq M_\varepsilon \varepsilon_n\right] P_{\theta_0}^{(n)}\left[d(\theta_0, \hat{\theta}_n) \leq M_\varepsilon \varepsilon_n\right] \\ &\geq P_{\theta_0}^{(n)}\left[r_\alpha > \frac{M_\varepsilon \varepsilon_n}{L_n}\right] (1 - \varepsilon) \end{aligned}$$

denote  $\rho_n = M_\varepsilon/L_n$ , then by definition of  $r_\alpha$ ,

$$P_{\theta_0}^{(n)}\left(\theta_0 \in \widehat{C}(L_n, \alpha)\right) \geq (1 - \varepsilon) P_{\theta_0}^{(n)}\left(\Pi\left(d(\theta, \hat{\theta}_n) \leq \rho_n \varepsilon_n \mid \mathbf{Y}\right) \leq 1 - \alpha\right).$$

In view of Lemma 3, if  $\rho_n \leq \delta/\sqrt{\log n}$  with  $\delta = \delta(\varepsilon, \alpha) > 0$  small enough and  $n$  large enough,

$$E_{\theta_0}^{(n)}\left(\pi\left(d(\theta, \hat{\theta}_n) \leq \rho_n \varepsilon_n \mid \mathbf{Y}\right)\right) \leq \varepsilon \alpha / (1 - \varepsilon).$$

Therefore for all  $\alpha > 0$ , and all  $\varepsilon > 0$  by Markov's inequality there exists  $L_{\varepsilon, \alpha} > 0$  large enough such that (16) holds.  $\square$

In the following section we prove Lemma 3.

4.1.1. *Proof of Lemma 3.* Let  $\hat{b}_n(k) = \min\{d^2(\theta, \hat{\theta}_n), \theta \in \Theta(k)\}$  and set  $\varepsilon_n = \varepsilon_n(k_n)$  and  $\Theta_n = \cup_k \Theta_n(k)$ . Note that on the event  $\{d(\theta_0, \hat{\theta}_n) \leq \varepsilon_n\}$ , if  $b(k) \geq 4\varepsilon_n^2$ ,  $\sqrt{\hat{b}_n(k)} \geq \sqrt{b(k)} - \varepsilon_n \geq \varepsilon_n$ . We thus have that  $\hat{b}_n(k) \leq \rho_n^2 \varepsilon_n^2$  only if  $b(k) \leq 4\varepsilon_n^2$ . This leads to

$$\begin{aligned} &\pi\left(\{d(\theta, \hat{\theta}_n) \leq \rho_n \varepsilon_n\} \cap \Theta_n \mid \mathbf{Y}\right) \\ &\leq \sum_k \mathbf{1}_{\hat{b}_n(k) \leq \rho_n^2 \varepsilon_n^2} \pi|_k\left(\{d(\theta, \hat{\theta}_n) \leq \rho_n \varepsilon_n\} \cap \Theta_n(k) \mid \mathbf{Y}, k\right) \pi_k(k \mid \mathbf{Y}) \\ &\leq \sum_k \mathbf{1}_{b(k) \leq 4\varepsilon_n^2} \mathbf{1}_{k \in \mathcal{K}_n(M)} \pi|_k\left(\{d(\theta, \hat{\theta}_n) \leq \rho_n \varepsilon_n\} \cap \Theta_n(k) \mid \mathbf{Y}, k\right) \pi_k(k \mid \mathbf{Y}) + o_{P_{\theta_0}}(1) \end{aligned}$$

for all  $\theta_0 \in \Theta_0$ , where the last equality follows from Lemma 1, for  $M$  satisfying (14).

We now study for all  $k \in \mathcal{K}_n(M)$ ,

$$(28) \quad \pi_{n,k} = \pi|_k\left(\{d(\theta, \hat{\theta}_n) \leq \rho_n \varepsilon_n\} \cap \Theta_n(k) \mid \mathbf{Y}\right).$$

In the set  $\{\theta \in \Theta_n(k) : d(\theta, \hat{\theta}_n) \leq \rho_n \varepsilon_n\}$  let  $\hat{\theta}_{[k]}$  satisfy  $d(\theta, \hat{\theta}_{[k]}) \leq \hat{b}_n(k)^{1/2} + \rho_n \varepsilon_n$  and introduce the notations

$$(29) \quad \bar{\Omega}_n(C) = \left\{ \max_{k \in \mathcal{K}_n(M)} e^{Ck} \frac{\int_{\Theta(k)} e^{\ell_n(\theta) - \ell_n(\theta_{[k]}^o)} \pi_{|k}(d\theta)}{\pi_{|k}(d(\theta, \theta_{[k]}^o)^2 \leq k/n)} \geq 1 \right\},$$

$$(30) \quad \Gamma_n(B) = \left\{ \max_{k \in \mathcal{K}_n(M)} \sup_{\Theta_n(k) \cap B_k(\theta_{[k]}^o, (M_0+1)\rho_n \varepsilon_n(k_n))} (\ell_n(\theta) - \ell_n(\theta_{[k]}^o) - Bk) < 0 \right\}.$$

In view of assumption **A4** we have  $P_{\theta_0}^{(n)}(\Gamma_n(B)) \rightarrow 1$  for large enough choice of  $B$  (depending on  $M$ ).

Note that following from Condition **A2(k)** for  $k \in \mathcal{K}_n(M)$  and by using the standard technique for lower bound for the likelihood ratio (e.g. Lemma 10 of [18]) we have, for any  $k \in \mathcal{K}_n(M)$ , with  $P_{\theta_0}^{(n)}$ -probability bounded from below by  $1 - \varepsilon^2/k$

$$(31) \quad \begin{aligned} \int_{\Theta(k)} e^{\ell_n(\theta) - \ell_n(\theta_{[k]}^o)} \pi_{|k}(d\theta) &\geq e^{-(\kappa+1/\varepsilon)k} \pi_{|k}(S_n(k, \kappa, \kappa')) \\ &\geq e^{-(\kappa+1/\varepsilon)k} \pi_{|k}(B_k(\theta_{[k]}^o, \sqrt{k/n})), \end{aligned}$$

hence  $P_{\theta_0}^{(n)}(\bar{\Omega}_n(\kappa + 1/\varepsilon)) \gtrsim 1 - \varepsilon^2$ .

Then we have, on  $\bar{\Omega}_n(\kappa + 1/\varepsilon) \cap \Gamma_n(B)$ , that for any  $k \in \mathcal{K}_n(M)$

$$(32) \quad \begin{aligned} \pi_{n,k} &\leq e^{(\kappa+B+1/\varepsilon)k} \frac{\pi_{|k}(\Theta_n(k) \cap \{d(\theta, \hat{\theta}_{[k]}) \leq 2\rho_n \varepsilon_n\})}{\pi_{|k}(d(\theta, \theta_{[k]}^o) \leq \sqrt{k/n})} \\ &\leq e^{(\kappa+B+1/\varepsilon)k} \frac{C_\infty^k \text{Vol}(B_k(\hat{\theta}_{[k]}, 2\rho_n \varepsilon_n))}{C_B^k \text{Vol}(B_k(\theta_{[k]}^o, \sqrt{k/n}))}. \end{aligned}$$

We recall that Remark 1 implies that  $k \asymp k_n$  for all  $k \in \mathcal{K}_n(M)$  and by definition of  $\varepsilon_n(k_n)$ ,  $n\varepsilon_n^2 \leq 2k_n \log n$ . Therefore if  $\rho_n \sqrt{\log n} \leq \delta$  we have  $\rho_n \varepsilon_n \leq 2\delta \sqrt{k_n/n} \lesssim \delta \sqrt{k/n}$  for all  $k \in \mathcal{K}_n(M)$  and hence by **A7(k)** (for all  $k \in \mathcal{K}_n(M)$ ) for small enough choice of the constant  $\delta > 0$  the right hand side of the preceding display is bounded by  $e^{-c(\delta)k_n}$ , concluding the proof of Theorem 3.

4.2. *Proof of Lemma 1.* We have

$$\pi_k(k|\mathbf{Y}) = \frac{\pi_k(k) \int_{\Theta(k)} e^{\ell_n(\theta) - \ell_n(\theta_0)} \pi_{|k}(d\theta)}{\sum_{k'} \pi_k(k') \int_{\Theta(k')} e^{\ell_n(\theta) - \ell_n(\theta_0)} \pi_{|k'}(d\theta)}.$$

Let  $m_n(k) = \int_{\Theta(k)} e^{\ell_n(\theta) - \ell_n(\theta_0)} \pi_k(d\theta)$  and  $\varepsilon_n = \varepsilon_n(k_n)$ . Next we give bounds for the marginal likelihood function, starting with a lower bound for  $m_n(k_n)$ . Let

$$(33) \quad \Omega_{n,0} = \left\{ m_n(k_n) > e^{-(\kappa + \kappa_1 + 1)k_n \log n} \right\}$$

and similarly to (31) we get following **A1** that  $P_{\theta_0}^{(n)}(\Omega_{n,0}^c) = o(1)$ . Furthermore, we show below that for every  $C, A > 0$  there exists a large enough choice of  $M$  such that

$$(34) \quad P_{\theta_0}^{(n)} \left( \int_{\Theta(k)} e^{\ell_n(\theta) - \ell_n(\theta_0)} d\pi_k(\theta) > e^{-Cn\varepsilon_n^2} \right) \leq 4e^{-n\varepsilon_n^2}$$

for all  $k \notin \mathcal{K}_n(M)$ ,  $k \leq Ak_n \log n$ .

Let  $\Omega_n(k) = \{m_n(k) \leq e^{-Cn\varepsilon_n^2}\}$ , and

$$\Omega_n = \Omega_{n,0} \cap_{k < Ak_n \log n; k \notin \mathcal{K}_n(M)} \Omega_n(k).$$

We have  $P_{\theta_0}^{(n)}(\Omega_n^c) = o(1)$  since  $k_n \log n = o(e^{n\varepsilon_n^2})$  and on  $\Omega_n$ ,

$$\begin{aligned} \pi_k(\{k < Ak_n \log n\} \cap \mathcal{K}_n(M)^c | \mathbf{Y}) &\leq \sum_{k < Ak_n \log n} \mathbf{1}_{k \in \mathcal{K}_n(M)^c} \frac{e^{-Cn\varepsilon_n^2} \pi_k(k)}{\pi_k(k_n) m_n(k_n)} \\ &\leq e^{-Cn\varepsilon_n^2 + c_2 k_n t(k_n) + (\kappa + \kappa_1 + 1)k_n \log n} \\ &\leq e^{-(C - \kappa - \kappa_1 - 1)k_n \log n + c_2 k_n t(k_n)}. \end{aligned}$$

In the case where  $t(k) = 1$  then choosing  $C > \kappa + \kappa_1 + 1$  leads to  $\pi_k(k \notin \mathcal{K}_n(M) : k < Ak_n \log n | \mathbf{Y}) = o(1)$ , while in the case  $t(k) = \log k$  the same result holds with  $C > \kappa + \kappa_1 + 1 + c_2$ . For  $k \notin \mathcal{K}_n(M)$ ,  $k \geq Ak_n \log n$  we also obtain using  $\Omega_{n,0}$  that

$$\begin{aligned} E_{\theta_0}^{(n)}(\pi_k(k \geq Ak_n \log n | \mathbf{Y})) &\leq e^{(\kappa + \kappa_1 + 1)k_n \log n + c_2 k_n t(k_n)} E_{\theta_0}^{(n)} \left( \sum_{k \geq Ak_n \log n} \pi_k(k) m_n(k) \right) + P_{\theta_0}^{(n)}(\Omega_{n,0}^c) \\ &\leq \pi_k(k \geq Ak_n \log n) e^{(\kappa + \kappa_1 + 1)k_n \log n + c_2 k_n t(k_n)} + o(1) \\ &\leq e^{-(c_1 A \log n - c_2)k_n t(Ak_n \log n) + (\kappa + \kappa_1 + 1)k_n \log n} + o(1) = o(1) \end{aligned}$$

as soon as

$$A > \frac{1}{c_1} \left( \frac{\kappa + \kappa_1 + 1}{t(k_n)} + \frac{c_2}{\log n} \right).$$



It remained to verify (34) for all  $k \notin \mathcal{K}_n(M)$ ,  $k < Ak_n \log n$ . If  $k < k_n$  we have for all  $\theta \in \Theta(k)$  that  $d^2(\theta_0, \theta) \geq b(k)$  and  $b(k) > k \log n/n$ , hence  $d^2(\theta_0, \theta) \geq \varepsilon_n^2(k)/2$ . If  $Ak_n \log n > k \geq k_n$  then  $k \geq M^2 k_n/2$  following from Remark 1. By slightly abusing our notation, consider slices  $\Theta_j(k) = \{j\varepsilon_n(k)/\sqrt{2} \leq d(\theta_0, \theta) \leq (j+1)\varepsilon_n(k)/\sqrt{2}\}$ ,  $j \geq 1$  of  $\Theta(k)$  and a coverage of the slice  $\Theta_j(k)$  into  $N_{n,j}(k)$  balls with centers  $\{\theta_{j,i}, i \leq N_{n,j}(k)\}$ . Note that for  $k < k_n$  we have  $b(k) \geq \varepsilon_n^2(k)/2$ , hence  $\Theta(k) = \cup_{j \geq 1} \Theta_j(k)$ . Next for each  $\theta_{j,i}$  consider the individual test  $\phi_n(j, i)$  defined in assumption **A5(k)**, and construct  $\phi_n(k) = \max_{j \geq J_0(k)} \max_{i=1}^{N_{n,j}(k)} \phi_n(j, i)$  where  $J_0(k) = 1$  if  $k < k_n$  and  $J_0(k) = J_0$  large enough if  $k \geq k_n$ . Each individual test satisfies

$$E_{\theta_0}^{(n)}(\phi_n(j, i)) \leq e^{-c_0 n j^2 \varepsilon_n^2(k)/2}, \quad \sup_{d(\theta_{j,i}, \theta) \leq \zeta j \varepsilon_n(k)/\sqrt{2}} E_{\theta}^{(n)}(1 - \phi_n(j, i)) \leq e^{-c_0 n j^2 \varepsilon_n^2(k)/2}.$$

Next note that for  $k < k_n$  and  $k \notin \mathcal{K}_n(M)$  we have  $2b(k) \geq \varepsilon_n^2(k) \geq M^2 k_n \log n/n > M^2 k \log n/n$  which by choosing  $M^2 \geq 2u_0^2$  implies that  $b(k) \geq u_0^2 k \log n/n$ . Therefore, in view of assumption **A5(k)**

$$E_{\theta_0}^{(n)}(\phi_n(k)) \leq \sum_{j \geq 1} e^{-c_0 n j^2 \varepsilon_n^2(k)/4}, \quad \sup_{\theta \in \Theta_n(k)} E_{\theta}^{(n)}(1 - \phi_n(k)) \leq e^{-c_0 n \varepsilon_n^2(k)/2}.$$

If  $k > k_n$  then  $\varepsilon_n^2(k) \geq k \log n/n$  and choosing  $J_0 \geq u_0 \vee 1$ , assumption **A5(k)** implies that

$$E_{\theta_0}^{(n)}(\phi_n(k)) \leq \sum_{j \geq J_0} e^{-c_0 n j^2 \varepsilon_n^2(k)/4} \lesssim e^{-c_0 J_0^2 k \log n/4},$$

$$\sup_{\theta \in \Theta_n(k) \cap B_k^c(\theta_0, J_0 \varepsilon_n(k))} E_{\theta}^{(n)}(1 - \phi_n(k)) \leq e^{-c_0 J_0^2 n \varepsilon_n^2(k)/2}.$$

Next choose  $M^2 \geq 8(C+1)/c_0$  so that

$$c_0 \varepsilon_n^2(k)/4 \geq M^2 c_0 \varepsilon_n^2(k_n)/8 \geq (C+1) \varepsilon_n^2(k_n) \quad \text{and} \quad c_0 J_0^2 \varepsilon_n^2(k)/4 \geq (C+1) \varepsilon_n^2(k_n).$$

Then for  $k < k_n$  following from assumption **A3(k)**,

$$(35) \quad P_{\theta_0}^{(n)} \left( \int_{\Theta(k)} e^{\ell_n(\theta) - \ell_n(\theta_0)} \pi_{|k}(d\theta) > e^{-Cn \varepsilon_n^2(k_n)} \right)$$

$$\leq E_{\theta_0}^{(n)}(\phi_n(k)) + e^{Cn \varepsilon_n^2(k_n)} \pi_{|k}(\Theta_n(k)^c) + e^{Cn \varepsilon_n^2(k_n)} \int_{\Theta_n(k)} E_{\theta}^{(n)}(1 - \phi_n(k)) \pi_{|k}(d\theta)$$

$$\leq 3e^{-n \varepsilon_n^2(k_n)}.$$

If  $k > k_n$ , following from assumptions **A3(k)**, **A5(k)** and **A6(k)**, and by applying Markov's inequality

$$\begin{aligned}
 (36) \quad & P_{\theta_0}^{(n)} \left( \int_{\Theta(k)} e^{\ell_n(\theta) - \ell_n(\theta_0)} \pi_{|k}(d\theta) > e^{-Cn\varepsilon_n^2(k_n)} \right) \\
 & \leq E_{\theta_0}^{(n)}(\phi_n(k)) + e^{Cn\varepsilon_n^2(k_n)} \pi_{|k}(\Theta_n(k) \cap B_k(\theta_0, J_0\varepsilon_n(k))) + \\
 & \quad + e^{Cn\varepsilon_n^2(k_n)} \pi_{|k}(\Theta_n(k)^c) + e^{Cn\varepsilon_n^2(k_n)} \int_{B_k(\theta_0, J_0\varepsilon_n(k))^c} E_{\theta}^{(n)}(1 - \phi_n(k)) \pi_{|k}(d\theta) \\
 & \leq 3e^{-n\varepsilon_n^2(k_n)} + e^{Cn\varepsilon_n^2(k_n) - (C+2)n\varepsilon_n^2(k_n) + k_n \log(C_\infty 2\pi)} \leq 4e^{-n\varepsilon_n^2(k_n)}
 \end{aligned}$$

for  $n$  large enough, terminating the proof of (34) and hence the proof of Lemma 1.

4.3. *Proof of Lemma 2.* Let  $M$  be as in Lemma 1, then

$$E_{\theta_0}^{(n)}(\pi(d(\theta, \theta_0) > M_1\varepsilon_n(k_n)|\mathbf{Y})) = E_{\theta_0}^{(n)}(\pi(\{d(\theta, \theta_0) > M_1\varepsilon_n(k_n)\} \cap \{k \in \mathcal{K}_n(M)\}|\mathbf{Y})) + o(1)$$

uniformly over  $\Theta_0$ . Let  $k \in \mathcal{K}_n(M)$ , then if  $k < k_n$  (assuming  $k_n > 1$ ), we have following Remark 1

$$\varepsilon_n^2(k) \geq b(k) \geq \frac{k \log n}{n} \geq \frac{k_n \log n}{R_0^{m+1}n} \geq \frac{\varepsilon_n^2(k_n)}{2R_0^{m+1}}$$

and if  $k > k_n$

$$\varepsilon_n^2(k) \geq \frac{k \log n}{n} > \frac{k_n \log n}{n} \geq \frac{\varepsilon_n^2(k_n)}{2}$$

so that

$$\frac{\varepsilon_n^2(k_n)}{2R_0^{m+1}} \leq \varepsilon_n^2(k) \leq M^2\varepsilon_n^2(k_n)$$

for all  $k \in \mathcal{K}_n(M)$ .

As in the proof of Lemma 1 define the tests  $\phi_n(j, i)$  with  $j \geq M_1$

$$\phi_n = \max_{k \in \mathcal{K}_n(M)} \max_{j \geq M_1} \max_{i \leq N_j(k)} \phi_n(j, i), \quad N_j(k) \leq \exp(c_0 n j^2 \varepsilon_n^2(k)/4).$$

We have,

$$E_{\theta_0}^{(n)}(\phi_n) \leq \sum_{k \in \mathcal{K}_n(M)} \sum_{j \geq M_1} e^{-c_0 n j^2 \varepsilon_n^2(k)/4} \leq 4M^2 k_n e^{-c_0 n M_1^2 \varepsilon_n^2(k_n)/(8R_0^{m+1})}$$

and for  $\theta \notin B_k(\theta_0, M_1 \varepsilon_n(k_n))$  with  $k \in \mathcal{K}_n(M)$

$$E_{\theta}^{(n)}(1 - \phi_n) \leq e^{-c_0 n M_1^2 \varepsilon_n^2(k_n)/(4R_0^{m+1})}.$$

Then,

$$\begin{aligned} E_{\theta_0}^{(n)}(\pi(d(\theta, \theta_0) > M_1 \varepsilon_n(k_n) | \mathbf{Y})) &\leq P_{\theta_0}^{(n)}(\Omega_{n,0}^c) + E_{\theta_0}^{(n)}(\phi_n) + o(1) \\ &\quad + e^{(\kappa + \kappa_1 + 1)k_n \log n + c_2 t(k_n)k_n - c_0 n M_1^2 \varepsilon_n^2(k_n)/(8R_0^{m+1})} = o(1) \end{aligned}$$

choosing  $M_1$  large enough.

4.4. *Proof of Theorem 2.* First we show that

$$(37) \quad P_{\theta_0}^{(n)}(\hat{k}_n \in \mathcal{K}_n(M)) \rightarrow 1.$$

In view of (35) and (36)

$$\begin{aligned} P_{\theta_0}^{(n)}\left(\sup_{k \notin \mathcal{K}_n(M), k \leq k_n} \int_{\Theta(k)} e^{\ell_n(\theta) - \ell_n(\theta_0)} \pi_{|k}(d\theta) > e^{-Cn\varepsilon_n^2(k_n)}\right) &\leq 3k_n e^{-n\varepsilon_n^2(k_n)}, \\ P_{\theta_0}^{(n)}\left(\sup_{k \notin \mathcal{K}_n(M), k \geq k_n} \int_{\Theta(k)} e^{\ell_n(\theta) - \ell_n(\theta_0)} \pi_{|k}(d\theta) > e^{-Cn\varepsilon_n^2(k_n)}\right) &\leq 4n^a e^{-k_n \log n}, \end{aligned}$$

where both terms on the right hand side tend to zero for  $k_n > a$ . Furthermore, from the argument below (33)

$$\int_{\Theta(k_n)} e^{\ell_n(\theta) - \ell_n(\theta_0)} \pi_{k_n}(d\theta) \geq e^{-(\kappa + \kappa_1 + 1)k_n \log n}$$

we get that

$$P_{\theta_0}^{(n)}\left(\sup_{k \notin \mathcal{K}_n(M)} m_n(k) < m_n(k_n)\right) \rightarrow 1,$$

leading to (37).

Next by using the notation (28) and following from (32) and (37) we have with  $P_{\theta_0}^{(n)}$ -probability tending to one that

$$\pi_{|\hat{k}}\left(\{d(\theta, \hat{\theta}_n) \leq \rho_n \varepsilon_n\} \cap \Theta_n | \mathbf{Y}\right) \leq \sum_{k \in \mathcal{K}_n(M)} \pi_{n,k} \lesssim k_n e^{-c(\delta)k_n} = o(1),$$

Then the proof of the first statement automatically follows from the proof of Theorem 1. The proof of the second statement follows by similar lines of reasoning as above combined with the proof of Lemma 2.

## APPENDIX A: PROOF OF THE PROPOSITIONS

**A.1. Proof of Proposition 1.** The proof of the first assertion is a consequence of Theorem 1, hence it is sufficient to verify the corresponding conditions, but before that we introduce some additional notations. Along the lines we use the notation  $\Phi_k(x_i) = (\phi_1(x_i), \dots, \phi_k(x_i))$ ,  $p_f$  for the corresponding density to the regression function  $f$  and  $\theta_{[k]}^o$  the  $d_n(\cdot, \cdot)$ -projection of  $\theta_0$  to  $\Theta(k) = \mathbb{R}^k$ . Furthermore, note that for  $\theta_1, \theta_2 \in \mathbb{R}^k$  we have that

$$d_n^2(f_{\theta_1}, f_{\theta_2}) = (\theta_1 - \theta_2)^T \left[ \frac{1}{n} \Phi_k^T \Phi_k \right] (\theta_1 - \theta_2).$$

Finally we note that the empirical hellinger distance  $h_n(\cdot, \cdot)$  between the densities indexed by the functions  $f_0$  and  $f_\theta$  with  $\theta \in \mathbb{R}^k$  is

$$h_n(p_{f_0}, p_{f_\theta}) = \frac{1}{n} \sum_{i=1}^n h(p_{f_0}(x_i), p_{f_\theta}(x_i)) = \frac{1}{n} \sum_{i=1}^n 1 - e^{-\frac{1}{4\sigma^2} (f_0(x_i) - \Phi_k(x_i)\theta)^2}.$$

Assuming that  $(f_0(x_i) - \Phi_k(x_i)\theta_{[k]}^o)^2 + (\theta_{[k]}^o - \theta)^T \Phi_k(x_i)^T \Phi_k(x_i) (\theta_{[k]}^o - \theta) \leq 1/2$  we can give lower and upper bounds for the right hand side of the preceding display as

$$(38) \quad h_n^2(p_{f_0}, p_{f_\theta}) \asymp d_n^2(f_0, f_\theta) \quad \text{and} \quad h_n^2(p_{f_{\theta_{[k]}^o}}, p_{f_\theta}) \asymp d_n^2(f_{\theta_{[k]}^o}, f_\theta).$$

Similarly the KL divergence between the densities indexed by  $f_0$  and  $\theta^T \Phi_k$  is

$$(39) \quad KL(p_{f_0}, p_{f_\theta}) = V_{2,0}(p_{f_0}, p_{f_\theta}) \asymp n d_n^2(f_{\theta_0}, f_\theta).$$

Let us define the sieve  $\Theta_n(k)$  as

$$\Theta_n(k) \equiv \{\theta \in \mathbb{R}^k : \|\theta\|_2 \leq n^Q\},$$

for some  $Q > 0$ .

Furthermore, let us introduce the notation  $\|f\|_n^2 = d_n(f, f)$  for  $f \in \mathbb{R}^n$ . Then we note that for every  $\theta \in \Theta(k)$

$$\|\mathbf{Y} - \Phi_k \theta\|_n^2 - \|\mathbf{Y} - \Phi_k \theta_{[k]}^o\|_n^2 = \|\Phi_k \theta_{[k]}^o - \Phi_k \theta\|_n^2 + 2d_n(\mathbf{Y} - \Phi_k \theta_{[k]}^o, \Phi_k(\theta_{[k]}^o - \theta)).$$

Besides,  $\mathbf{Y} = f_{0,n} + \mathbf{Z}$  (where  $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)$ ) and hence the  $P_{f_0}^{(n)}$ -expected value of the second term on the right hand is zero following from

the orthonormality of the  $d_n(\cdot, \cdot)$ -projection of  $f_0$  into the sub-space  $\{\Phi_k \theta : \theta \in \Theta(k)\}$  and  $E_{f_0}^{(n)} \mathbf{Z} = 0$ . Therefore

$$(40) \quad E_{f_0}^{(n)} \log \frac{p_{\theta_{[k]}^o}}{p_\theta} = n d_n^2(f_{\theta_{[k]}^o}, f_\theta).$$

By Chebyshev's inequality we have that with  $P_{f_0}^{(n)}$ -probability tending to one

$$\begin{aligned} & \sup_{k \in \mathcal{K}_n(M)} \sup_{\theta \in B_k(\theta_{[k]}^o, M \rho_n \varepsilon_n(k))} \left| \ell_n(\theta_{[k]}^o) - \ell_n(\theta) E_{f_0}^{(n)} \log \frac{p_{f_{\theta_{[k]}^o}^{(n)}}}{p_{f_\theta}^{(n)}} \right| \\ & \leq \sup_{k \in \mathcal{K}_n(M)} \sup_{d_n(f_\theta, f_{\theta_{[k]}^o}) \leq M' \rho_n \varepsilon_n(k)} \sum_{i=1}^n |Z_i| |f_\theta(x_i) - f_{\theta_{[k]}^o}(x_i)| \\ & \leq \|\mathbf{Z}\|_2 \sup_{k \in \mathcal{K}_n(M)} \sup_{\theta \in B_k(\theta_{[k]}^o, M \rho_n \varepsilon_n(k))} d_n(f_\theta, f_{\theta_{[k]}^o}) \\ & \leq M_n \sqrt{n} \rho_n \sup_{k \in \mathcal{K}_n(M)} \varepsilon_n(k) \lesssim M_n \rho_n \sqrt{k_n} \sqrt{\log n} \lesssim k_n, \end{aligned}$$

for any sequence  $M_n$  tending to infinity (slow enough). Furthermore, following from (40)  $E_{f_0}^{(n)} |(\ell_n(\theta_{[k]}^o) - \ell_n(\theta))| \lesssim \rho_n^2 n \varepsilon_n(k_n)^2 \lesssim k_n$  for  $\rho_n = O(1/\sqrt{\log n})$ , providing us Condition **A4**.

From Lemma 2 of [1] follows that for all  $k \leq A k_n \log n$

$$\pi_{|k}(\Theta_n(k)^c) \lesssim e^{-\frac{s_0}{2} n^{Qp} \min((A k_n \log n)^{1-p/2}, 1)},$$

hence by large enough choice of  $Q$  it is bounded from above by  $e^{-c_2 n \varepsilon_n(k)^2}$ , for every constant  $c_2 > 0$ , hence assumption **A3(k)** holds for all  $k \leq A k_n \log n$ .

The testing part of condition **A5(k)** is verified in Corollary 2 on page 149 of [5]. For the entropy part of condition **A5(k)** we note that following from (21) we have that for every  $0 < \zeta < 1$  and  $c_0 > 0$  there exists  $u_0 > 0$  satisfying

$$N(\zeta u, \Theta_n(k), d_n(\cdot, \cdot)) \leq N((\zeta/C_0)u, \Theta_n(k), \|\cdot\|_2) \leq c_0 u_0^2 k \log n,$$

for every  $u \leq n^{-c_3}$  (for any  $c_3 > 0$ ).

Then to prove condition **A6(k)** for  $M^2 k_n/2 \leq k \leq A k_n \log n$  we note that following assumption (21)

$$\text{Vol}(B_k(\theta_0, J_0 \varepsilon_n(k))) \leq \text{Vol}(B_k(\theta_{[k]}^o, (C_0 + 1) J_0 \varepsilon_n(k), \|\cdot\|_2)) \leq e^{k \log(C(J_0 + 1) \varepsilon_n(k))},$$

for some universal constant  $C > 0$  not depending on  $k$ . Hence for  $k_n = O(n^\gamma)$  with  $\gamma < 1$  we have that the right hand side of the preceding display is bounded from above by  $e^{-C'k \log n} \leq e^{-(C'M^2/2)k_n \log n} \leq e^{-(C'M^2/4)n\varepsilon_n^2(k_n)}$  for  $k \geq (M^2/2)k_n$  and some constant  $C' > 0$  depending on  $\gamma$ . Therefore by sufficiently large choice of the constant  $M$  in the definition of  $\mathcal{K}_n(M)$  given below (7) condition **A6(k)** holds.

Next we deal with condition **A2(k)** for  $k \in \mathcal{K}_n(M)$ . Similarly to (40) we have  $V_{f_0}^{(n)}(\log \frac{p_{\theta_{[k]}^o}}{p_\theta}) = d_n^2(f_{\theta_{[k]}^o}, f_\theta)$ , resulting  $B_k(\theta_{[k]}^o, \sqrt{k/n}) \subset \mathcal{S}_n(k, 1, 1)$ . Furthermore, by (21)

$$\text{Vol}[B_k(\theta_{[k]}^o, \sqrt{k/n}, d_n(\cdot, \cdot))] \geq \text{Vol}[B_k(\theta_{[k]}^o, \sqrt{k/n}/C_0, \|\cdot\|_2)] \geq e^{-\kappa_1 k \log n},$$

for some large enough  $\kappa_1 > 0$ , which combined with assumption **C2** provides us **A2(k)**. Condition **A1** follows along the same line of reasoning.

Finally, one can easily see that the small ball volume condition **A7(k)** is satisfied:

$$(41) \quad \frac{\text{Vol}(B_k(\tilde{\theta}, \delta\sqrt{k/n}, d_n(\cdot, \cdot)))}{\text{Vol}(B_k(\theta_{[k]}^o, \sqrt{k/n}, d_n(\cdot, \cdot)))} \leq \frac{\text{Vol}(B_k(\tilde{\theta}, C_0\delta\sqrt{k/n}, \|\cdot\|_2))}{\text{Vol}(B_k(\theta_{[k]}^o, C_0^{-1}\sqrt{k/n}, \|\cdot\|_2))} = \left(\frac{C_0\delta}{C_0^{-1}}\right)^k = e^{k \log(C_0^2\delta)}.$$

It remained to deal with the second assertion of the proposition. In view of Corollary 1 it is sufficient to give an upper bound for  $\varepsilon_n(k_n)$ . Let us introduce first the notation  $\theta_{0,[k]} = (\theta_{0,1}, \theta_{0,2}, \dots, \theta_{0,k}) \in \mathbb{R}^k$ . Then for  $\theta_0 \in S^\beta(M)$  with  $\beta > \beta_0$  we have that  $d_n(f_{\theta_0}, f_{\theta_{0,[K_n]}}) \leq \|\Delta_{K_n}\|_\infty \lesssim \sum_{i=K_n+1}^\infty |\theta_{0,i}| \lesssim K_n^{-(\beta-1/2)}$ . Therefore by triangle inequality and assumption (21) we get that

$$(42) \quad \begin{aligned} b(k)^{1/2} &\leq d_n(f_{\theta_{0,[k]}}, f_{\theta_{0,[K_n]}}) + d_n(f_{\theta_0}, f_{\theta_{0,[K_n]}}) \\ &\lesssim \left(\sum_{i=k+1}^{K_n} \theta_{0,i}^2\right)^{1/2} + K_n^{-(\beta-1/2)} \lesssim k^{-\beta} + K_n^{-(\beta-1/2)}. \end{aligned}$$

Hence by taking  $\bar{k}_n = C(n/\log n)^{1/(1+2\beta)}$  (with large enough constant  $C > 0$ ) we get that  $b(\bar{k}_n) \leq \bar{k}_n \log n/n$  and as a consequence  $k_n \leq \bar{k}_n$ . In view of assumption  $K_n \geq n^{\frac{\beta_0}{(1+2\beta_0)(\beta_0-1/2)}}$  this leads to

$$\varepsilon_n^2(k_n) \leq 2k_n \log n/n \leq 2\bar{k}_n \log n/n \lesssim (n/\log n)^{-2\beta/(1+2\beta)}.$$

**A.2. Proof of Proposition 2.** The first assertion of the theorem follows from Theorems 1 and 2, hence it is sufficient to verify the conditions **A1-A7(k)**. As a first step in doing so we note that

$$(43) \quad \|p_0 - p_\theta\|_2^2 = \int_0^1 (p_0 - p_\theta)^2(x) dx \geq \int_0^1 (\sqrt{p_0} - \sqrt{p_\theta})^2(x) p_0(x) dx \geq c_0 h^2(p_0, p_\theta)$$

where  $c_0$  is the lower bound for the density  $p_0$ . Furthermore, for all  $k$

$$\inf_{\theta \in \mathbb{R}^k} \|p_0 - p_\theta\|_2^2 = \|p_0 - p_{\theta_{[k]}^o}\|_2^2$$

with  $\theta_{[k]}^o = (\int_{I_j} p_0(x) dx, j \leq k) \in \Theta(k)$  so that  $b(k) \lesssim \|p_0 - p_{\theta_{[k]}^o}\|_2^2$ . Moreover, set  $\eta_j^o = \int_{I_j} \sqrt{p_0}(x) dx \asymp 1/k$ , then

$$\begin{aligned} h^2(p_0, p_\theta) &\geq \sum_j \int_{I_j} (\sqrt{p_0} - k\eta_j^o)^2 dx = \sum_j \int_{I_j} \frac{(p_0 - k^2(\eta_j^o)^2)^2}{(\sqrt{p_0}(x) + k\eta_j^o)^2} dx \\ &\geq \frac{1}{2C_0} \sum_j \int_{I_j} (p_0 - k^2(\eta_j^o)^2)^2 dx \geq \frac{1}{2C_0} \|p_0 - p_{\theta_{[k]}^o}\|_2^2 \end{aligned}$$

and  $b(k) \asymp \|p_0 - p_{\theta_{[k]}^o}\|_2^2$ . First of all we note that since the density  $p_0$  is bounded from below and above by some positive constants, the Kullback-Leibler divergence and second moment of the likelihood ratio are both bounded by the square hellinger distance, providing us Condition **A1**, see Lemma 8.2 of [17]. Condition **A2(k)** with  $k \in \mathcal{K}_n(M)$  is easily verified for all  $k_n = o(\sqrt{n})$ . Indeed, using the notation  $\theta_j^o = \theta_{[k],j}^o$ ,

$$(44) \quad \frac{k}{n} \geq h^2(p_{\theta_{[k]}^o}, p_\theta) = \sum_j (\sqrt{\theta_j} - \sqrt{\theta_j^o})^2 = \sum_j \frac{(\theta_j - \theta_j^o)^2}{(\sqrt{\theta_j} + \sqrt{\theta_j^o})^2}.$$

More over for all  $j$ ,

$$\sqrt{\theta_j} \leq \sqrt{\theta_j^o} + |\sqrt{\theta_j} - \sqrt{\theta_j^o}| \leq \sqrt{C_0}/\sqrt{k} + \sqrt{k}/\sqrt{n} \leq \frac{2\sqrt{C_0}}{\sqrt{k}}$$

and similarly  $\theta_j \geq c_0/(2k)$ , so that

$$(45) \quad \sum_j (\theta_j - \theta_j^o)^2 \leq \frac{9C_0}{n},$$

which in turns implies by Taylor's series expansion that

$$(46) \quad \begin{aligned} \int p_0 \log \left( \frac{p_{\theta_{[k]}^o}}{p_\theta} \right) &= \sum_{j=1}^k \theta_j^o \log(\theta_j^o / \theta_j) \leq k c_0^{-1} \sum_{j=1}^k (\theta_j^o - \theta_j)^2 \leq \frac{9C_0 k}{c_0 n}, \\ \int p_0 \log^2 \left( \frac{p_{\theta_{[k]}^o}}{p_\theta} \right) &\leq \frac{9C_0 k}{c_0 n}. \end{aligned}$$

The second assertion follows for instance from Section A.4 of [36].

By taking  $\Theta_n(k) = \Theta(k)$  condition **A3(k)** automatically holds. Moreover, for all  $u$  such that  $u^2 \gtrsim k/n$  the entropy condition (10) is verified with  $d(\cdot, \cdot)$  the hellinger metric, see for instance the proof of Proposition 3.6 in [37]. This together with the existence of tests with respect of the hellinger distance (with  $\zeta = 1/18$ ), see [4], verifies condition **A5(k)**.

Condition **A7(k)** need to be verified for  $k \in \mathcal{K}_n(M)$  and is more involved because the centers of the balls can be different. We first bound from above  $\text{Vol}(B_k(\tilde{\theta}, \delta\sqrt{k/n}))$ . Let  $\tilde{\theta} \in \Theta(k)$  such that  $h^2(p_{\theta_{[k]}^o}, p_{\tilde{\theta}}) \leq M_0^2 \varepsilon_n^2$ , then similarly to (45) we can derive that  $k\|\theta_{[k]}^o - \tilde{\theta}\|_2^2 \lesssim \varepsilon_n^2 \lesssim k_n \log n/n$  and  $\tilde{\theta}_j \gtrsim 1/k$  for all  $j \leq k$ . We then, using the same computations as before, have that there exists  $C_1 > 0$  depending only on  $c_0, C_0$  such that  $h(p_{\tilde{\theta}}, p_\theta)^2 \geq k\|\theta - \tilde{\theta}\|_2^2 / C_1$  so that

$$(47) \quad \text{Vol}(B_k(\tilde{\theta}, \delta\sqrt{k/n})) \leq \frac{\pi^k \left( \delta\sqrt{C_1/n} \right)^{k-1}}{\Gamma(k/2 + 1)}$$

for some  $C_1 > 0$ . We now bound from below  $\text{Vol}(B_k(\theta_{[k]}^o, \sqrt{k/n}))$ . Note first that  $\theta_j^o \geq c_0/k$  for all  $j \leq k$  so that similarly to (43) we get that  $h^2(p_{\theta_{[k]}^o}, p_\theta) \leq \|\theta - \theta_{[k]}^o\|_2^2 k / c_0$ . Therefore

$$\text{Vol}(B_k(\theta_{[k]}^o, \sqrt{k/n})) \geq \frac{\pi^k \left( \sqrt{c_0/n} \right)^{k-1}}{\Gamma(k/2 + 1)}$$

and

$$\begin{aligned} &\log \text{Vol}(B_k(\tilde{\theta}, \delta\sqrt{k/n})) - \log \text{Vol}(B_k(\theta_{[k]}^o, \sqrt{k/n})) \\ &\leq (k-1) \log \left( \delta\sqrt{C_1/c_0} \right) \leq -\frac{(k-1) \log(1/\delta)}{2} \end{aligned}$$

as soon as  $\delta < c_0/(2C_1)$  and condition **A7(k)** is proved.

Next we note that by triangle inequality and  $h^2(p_0, p_{\theta_{[k]}^o}) \lesssim b(k) \leq \varepsilon_n^2(k)$  we have that there exists  $J_1 > 0$  such that  $B_k(\theta_0, J_0 \varepsilon_n(k)) \subset B_k(\theta_{[k]}^o, (J_0 +$



$J_1)\varepsilon_n(k))$ . Therefore in view of (47), there exists  $C > 0$  such that for every  $k \geq M^2/2$

$$\text{Vol}(B_k(\theta_0, J_0\varepsilon_n(k))) \leq e^{Ck-(k/2)\log n} \leq e^{-(M^2/4-C/\log n)k_n \log n},$$

hence condition **A6(k)** with  $k \geq M^2/2$  holds for appropriately large choice of the parameter  $M > 0$ .

Finally we verify **A4**. First we note that

$$\begin{aligned} \ell_n(\theta) - \ell_n(\theta_{[k]}^o) - E_{\theta_0}^{(n)}(\ell_n(\theta) - \ell_n(\theta_{[k]}^o)) \\ = \sum_{j=1}^k (n_j - n\theta_j^o)(\log \theta_j - \log \theta_j^o) \\ = \sum_{j=1}^k (n_j - n\theta_j^o) \left( \frac{\theta_j - \theta_j^o}{\theta_{0,j}} - \frac{(\theta_j - \theta_j^o)^2}{\theta_j^2} \right), \end{aligned}$$

for some  $\bar{\theta}_j \in [\theta_j, \theta_j^o] \cup [\theta_j^o, \theta_j]$ . By Cauchy-Schwartz we get that

$$\begin{aligned} \sup_{B_k(\theta_{[k]}^o, M\rho_n\varepsilon_n(k))} \sum_{j=1}^k (n_j - n\theta_j^o) \frac{\theta_j - \theta_j^o}{\theta_j^o} \\ \leq \sup_{B_k(\theta_{[k]}^o, M\rho_n\varepsilon_n(k))} \sqrt{\sum_{j=1}^k (n_j - n\theta_j^o)^2} \sqrt{\sum_{j=1}^k \frac{(\theta_j - \theta_j^o)^2}{\theta_{0,j}^2}} \\ (48) \quad \leq \sqrt{k\rho_n^2 \log n/n} \sqrt{\sum_{j=1}^k (n_j - n\theta_j^o)^2}, \end{aligned}$$

where the last inequality follows from (44). Then we note that by the properties of the categorical random variable

$$\sum_{j=1}^k E_{p_0}^{(n)}(n_j - n\theta_j^o)^2 = n \sum_{j=1}^k \theta_j^o(1 - \theta_j^o) \leq n.$$

Moreover, straightforward but tedious computation implies that

$$V_{p_0}^{(n)} \left( \sum_{j=1}^k (n_j - n\theta_j^o)^2 \right) \lesssim \frac{n^2}{k}.$$

Hence for  $k \in \mathcal{K}_n(M)$ , using a Chebyshev inequality, if  $B_n > 1$

$$P_0 \left( \sum_{j=1}^k (n_j - n\theta_j^o)^2 > B_n n \right) \lesssim \frac{n^2}{k(B_n - 1)^2 n^2} \lesssim \frac{1}{k(B_n - 1)^2}.$$

Therefore by combining the preceding display with (48) taking  $\rho_n = O(1/\sqrt{\log n})$  and with (46) and by taking large enough  $B_n$  we prove that **A4** holds with  $o(1)$  replaced by arbitrary small  $\varepsilon > 0$ . To achieve  $o(1)$  we have to take  $\rho_n = o(1/\sqrt{\log n})$  and  $B_n = o(\log^{-1/2} n \rho_n^{-1})$ .

### A.3. Proof of Proposition 3.

A.3.1. *Proof.* To prove the first assertion it is sufficient to verify conditions **A1-A7(k)**. Before that we note that we take  $d(\cdot, \cdot)$  to be the hellinger distance  $h(\cdot, \cdot)$  and we choose

$$\Theta_n(k) = \{\theta \in \mathbb{R}^k; \|\theta\|_2 \leq R_n(k)\}, \quad R_n(k) = R_0(n\varepsilon_n(k)^2)^{1/p},$$

for some large enough  $R_0 > 0$ . Define  $\theta_{[k]}^o$  to be the Kullback-Leibler projection of  $\theta_0$  onto  $\Theta(k)$  which exists and is unique by convexity of  $\theta \rightarrow KL(\theta_0, \theta)$ . Denote also  $\theta_{0,[k]} = (\theta_{0,1}, \dots, \theta_{0,k})$ .

To prove **A5'(k)** for  $k \notin \mathcal{K}_n(M)$ ,  $k \leq Ak_n \log n$  we introduce the sets

$$(49) \quad B_{n,j}(k) = \Theta_n(k) \cap \{j\varepsilon_n(k_n) \leq \|\theta - \theta_{0,[k]}\|_2 \leq (j+1)\varepsilon_n(k_n)\}.$$

Let  $\theta \in B_{n,j}(k)$ , then  $\|\theta\|_1 \leq \|\theta_{0,[k]}\|_1 + \|\theta - \theta_{0,[k]}\|_1 \leq \|\theta_0\|_1 + \sqrt{k}(j+1)\varepsilon_n(k_n) \lesssim 1 + (j+1)\sqrt{k}\sqrt{k_n \log n/n}$ , so that Lemma 4 implies that if  $j \leq J_n := J_1\sqrt{n}/\sqrt{k k_n \log n}$ , with arbitrary  $J_1 > 0$ ,  $\|\theta - \theta_0\|_2 \asymp h(f_0, f_\theta)$  and  $h(f_0, f_\theta) \gtrsim j\varepsilon_n(k)$ . Note also that for  $k < k_n$ ,  $h(f_0, f_\theta) \geq \sqrt{b(k)} \geq \varepsilon_n(k)/2$ . Hence for  $j \leq J_n$ , conditions (11) and (12) are verified with  $c(k, j) \geq (j+1)c$  for some  $c > 0$  (as in the proof of Proposition 3.3 in [37]). We then collapse the sets  $B_{n,j}(k)$ , with  $j > J_n$  into a single set denoted  $\bar{B}_{n, J_n+1}(k)$ . In  $j > J_n$  as in the proof of Proposition 3.3 in [37]

$$h(f_\theta, f_{\theta_{0,[k]}}) \gtrsim \|\theta - \theta_{0,[k]}\|_2 / \left( \sqrt{k}\varepsilon_n(k)j + |\log(j\varepsilon_n(k))| \right) \gtrsim 1/\sqrt{k}$$

and since  $f_0 \in \mathcal{S}^{\beta_0}$  with  $\beta_0 > 1/2$ ,

$$h(f_\theta, f_0) \geq c_1 k^{-1/2} - c_2 k^{-\beta_0} \geq c_1 k^{-1/2}/2 \geq c(k, j)\varepsilon_n(k)$$

for  $k > (2c_2/c_1)^{1/(\beta_0-1/2)}$  and  $c(k, j) \gtrsim k^{-1/2}\varepsilon_n(k)^{-1}$ . If  $k \leq (2c_2/c_1)^{1/(\beta_0-1/2)}$  and  $k \notin \mathcal{K}_n(M)$ ,  $h(f_0, f_\theta)^2 \geq b(k) \geq M^2 k_n \log n / (2n)$ .

Also, from equation (3.2) of [34], if  $\|\theta - \theta'\|_2 \leq n^{-1}/\sqrt{k}$ ,  $\theta, \theta' \in \Theta(k)$ ,

$$h(f_\theta, f_{\theta'}) \leq 4\|\theta - \theta'\|_1 \leq 4n^{-1}$$

and

$$N(4n^{-1}, \cup_{j>J_n} B_{n,j}(k), h(\cdot, \cdot)) \leq N(n^{-1}/\sqrt{k}, \Theta_n(k), \|\cdot\|_2) \leq (CR_n(k)\sqrt{k}/n)^k \leq e^{-Ck \log n}$$

for some  $C > 0$ . Since  $k \log n = o(n/k)$  condition (13) is verified. Finally condition (12) is verified, noting that for all  $k \leq Ak_n \log n$  with  $k \notin \mathcal{K}_n(M)$  and  $M$  large enough,

$$\sum_{j \leq J_n} e^{-c_1 c(k,j)^2 n \varepsilon_n^2(k)/2} + e^{-c_1 c(k, J_n+1)^2 n \varepsilon_n^2(k)/2} \leq \sum_{j \leq J_n} e^{-c_1 c^2 n j^2 \varepsilon_n^2(k)/2} + e^{-c_1 M^2 k_n \log n/2}.$$

Next we deal with condition **A4**. Let  $k \in \mathcal{K}_n(M)$ . Denote  $\Phi(\mathbf{Y}) = (\sum_{i=1}^n \phi_j(Y_i), j = 1, \dots, k)^T$ . By Cauchy-Schwarz inequality we get that

$$\begin{aligned} \left| \ell_n(\theta_{[k]}^o) - \ell_n(\theta) - n E_{f_0}^{(n)} \log \frac{f_{\theta_{[k]}^o}}{f_\theta} \right| &= \left| (\theta_{[k]}^o - \theta)^T (\Phi(\mathbf{Y}) - E_{f_0}^{(n)} \Phi(\mathbf{Y})) \right| \\ &\leq \|\theta - \theta_{[k]}^o\|_2 \|\Phi(\mathbf{Y}) - E_{f_0}^{(n)} \Phi(\mathbf{Y})\|_2. \end{aligned}$$

Furthermore, by noting that in view of Lemma 4 we have that  $h(f_\theta, f_{\theta_{[k]}^o}) \asymp \|\theta - \theta_{[k]}^o\|_2$  for all  $\theta \in B_k(\theta_{[k]}^o, \rho_n \varepsilon_n(k))$  and

$$\sup_{\theta \in \Theta_n(k) \cap B_k(\theta_{[k]}^o, \rho_n \varepsilon_n(k))} \|\theta - \theta_{[k]}^o\|_2 \lesssim \rho_n \varepsilon_n(k) \lesssim \rho_n \sqrt{\log n} \sqrt{k/n}.$$

Furthermore,

$$E_{f_0}^{(n)} \|\Phi(\mathbf{Y}) - E_{f_0}^{(n)} \Phi(\mathbf{Y})\|_2^2 \leq E_{f_0}^{(n)} \|\Phi(\mathbf{Y})\|_2^2 \leq kn \|f_{\theta_0}\|_\infty$$

we get by applying Markov's inequality that with  $P_{f_0}^{(n)}$ -probability larger than  $1 - \varepsilon$

$$\sup_{k \in \mathcal{K}_n} \sup_{\theta \in \Theta_n(k) \cap B_k(\theta_{[k]}^o, \rho_n \varepsilon_n(k))} \left| \ell_n(\theta_{[k]}^o) - \ell_n(\theta) - E_{f_0}^{(n)} \log \frac{f_{\theta_{[k]}^o}}{f_\theta} \right| \lesssim \varepsilon^{-1} k_n \rho_n \sqrt{\log n} \lesssim k_n,$$

for  $\rho_n = O(1/\sqrt{\log n})$ . Moreover in view of Lemma 4, for all  $B > 0$ ,

$$\begin{aligned} (50) \quad E_{f_0} \log \frac{f_{\theta_{[k]}^o}}{f_\theta} &= (\theta - \theta_{[k]}^o)^T E_{f_0}(\Phi(Y_1)) + \log E_{f_{\theta_{[k]}^o}}(e^{(\theta - \theta_{[k]}^o)^T \Phi(Y_1)}) \\ &\lesssim \|\theta - \theta_{[k]}^o\|_2^2 \lesssim h^2(f_{\theta_{[k]}^o}, f_\theta) \lesssim \rho_n \varepsilon_n^2(k) \leq (B/2)k/n \end{aligned}$$

as soon as  $h(f_{\theta_{[k]}^o}, f_\theta) \lesssim \rho_n \varepsilon_n(k)$  and  $\rho_n \varepsilon_n(k) \leq \delta B$  for some  $\delta > 0$  small enough, which terminates the proof of **A4**.

Then we note that condition **A3(k)** is verified in the proof of Condition (A) of [34]. Furthermore, we note that from the proof of Proposition 3.3 of [37] (and similarly to the above computations for  $\theta_{[k]}^o$ ) follows that for

$\theta \in B_k(\theta_0, \varepsilon_n(k_n))$  we have that  $h(f_0, f_\theta) \gtrsim \|\theta - \theta_0\|_2$  and therefore  $\|\theta - \theta_0\|_1 = O(1)$ . Hence in view of Lemma 5.a of [37] **A1** holds.

Condition **A2(k)** for  $k \in \mathcal{K}_n(M)$  is dealt with using (50), together with Lemma 4. Indeed (50) implies that

$$E_{f_0} \log \frac{f_{\theta_{[k]}^o}}{f_\theta} \lesssim \|\theta - \theta_{[k]}^o\|_2^2 \lesssim h^2(f_{\theta_{[k]}^o}, f_\theta) \leq Ck/n$$

for some  $C > 0$  and all  $\theta \in B_k(\theta_{[k]}^o, \sqrt{k/n})$ . Moreover we then also have

$$V_{f_0} \log \frac{f_{\theta_{[k]}^o}}{f_\theta} = nE_{f_0}[(\theta - \theta_{[k]}^o)^T \Phi(Y_1)]^2 \leq \|f_0\|_\infty \|\theta - \theta_{[k]}^o\|_2^2,$$

which terminates the first inequality of **A2(k)**. The second inequality is verified following the lower bound of the term  $P_1$  in the proof of Condition (C) of [34].

We now study condition **A6(k)**. We have that for every  $(M^2/2)k_n \leq k \leq Ak_n \log n$  and  $\theta \in B_k(\theta_0, J_0 \varepsilon_n(k))$  and all  $J_0 > 0$  fixed,

$$(51) \quad h(f_0, f_\theta) \asymp \|\theta - \theta_0\|_2, \quad \text{and} \quad \theta \in \Theta_n(k)$$

so that there exists  $C > 0$  with

$$\text{Vol}(\Theta_n(k) \cap B_k(\theta_0, J_0 \varepsilon_n(k))) \leq e^{-Ck \log n} \leq e^{-(CM^2/4)n\varepsilon_n^2(k_n)},$$

which is further bounded from above by  $e^{-C'n\varepsilon_n^2(k_n)}$  for arbitrary  $C' > 0$  by large enough choice of the parameter  $M$ , and **A6(k)** holds.

Finally we deal with condition **A7(k)** for  $k \in \mathcal{K}_n(M)$ . Similarly to the assertion before we get for every  $\tilde{\theta}, \theta \in B_k(\theta_{[k]}^o, M_0 \varepsilon_n(k_n))$  and  $k \in \mathcal{K}_n(M)$ ,  $\|\theta\|_1 \vee \|\tilde{\theta}\|_1 \leq \|\theta_{[k]}^o\|_1 + o(1)$  so that  $h(f_\theta, f_{\tilde{\theta}}) \asymp \|\tilde{\theta} - \theta\|_2$ . Hence there exist positive constants  $C_1$  and  $C_2$  such that

$$\begin{aligned} B_k(\tilde{\theta}, \delta \sqrt{k/n}) &\subset B_k(\tilde{\theta}, C_1 \delta \sqrt{k/n}, \|\cdot\|_2) \quad \text{and} \\ B_k(\theta_{[k]}^o, \sqrt{k/n}) &\supset B_k(\theta_{[k]}^o, C_2 \sqrt{k/n}, \|\cdot\|_2) \end{aligned}$$

and the statement follows from (41).

We finally verify the second statement of Proposition 3. Following Corollary 1 it is sufficient to show that for  $\theta_0 \in S^\beta(L)$  we have  $\varepsilon_n(k_n) \lesssim (n/\log n)^{-\beta/(1+2\beta)}$ . Note that

$$\begin{aligned} h^2(f_{\theta_0}, f_{\theta_{[k]}^o}) &\lesssim e^{c_1(\|\theta_0\|_1 \|\theta_0 - \theta_{[k]}^o\|_1)} \|\theta_0 - \theta_{[k]}^o\|_2^2 \\ &\lesssim e^{c_1(\|\theta_0\|_1 \|\theta_0 - \theta_{0,[k]}\|_1 + \|\theta_{0,[k]} - \theta_{[k]}^o\|_1)} (\|\theta_0 - \theta_{0,[k]}\|_2^2 + \|\theta_{0,[k]} - \theta_{[k]}^o\|_2^2) \\ &\lesssim k^{-2\beta} \sum_{i=k+1}^{\infty} \theta_{0,i}^2 i^{2\beta} \lesssim k^{-2\beta}. \end{aligned}$$

Hence by choosing  $\bar{k}_n = C(n/\log n)^{1/(1+2\beta)}$  we get that  $b(\bar{k}_n) < \bar{k}_n \log n/n$  for sufficiently large  $C > 0$  and therefore  $k_n \leq \bar{k}_n$ . We conclude the proof by noting that

$$\varepsilon_n(k_n) \leq 2\sqrt{k_n(\log n)/n} \leq 2\sqrt{\bar{k}_n(\log n)/n} \lesssim (n/\log n)^{-\beta/(1+2\beta)}.$$

### A.3.2. Technical Lemmas.

LEMMA 4. *For all  $A > 0$ , there exist  $C_A, c_1, c_2 > 0$  such that for any and all  $\theta_1, \theta_2 \in \ell_2 \cap \ell_1$ , if  $\|\theta_1 - \theta_2\|_2 \leq A$*

$$(52) \quad \begin{aligned} h^2(f_{\theta_1}, f_{\theta_2}) &\leq C_A e^{c_1(\|\theta_1\|_1 + \|\theta_1 - \theta_2\|_1)} \|\theta_1 - \theta_2\|_2^2, \\ h^2(f_{\theta_1}, f_{\theta_2}) &\geq C_A^{-1} e^{-c_2(\|\theta_1\|_1 + \|\theta_1 - \theta_2\|_1)} \|\theta_1 - \theta_2\|_2^2. \end{aligned}$$

Let  $\theta_{[k]}^o$  be the Kullback-Leibler projection of  $\theta_0$  onto  $\Theta(k)$ , then  $\theta_{[k]}^o$  satisfies

$$E_{f_0}(\phi_j) = E_{f_{\theta_{[k]}^o}}(\phi_j), \quad \forall j \leq k, \quad \text{and} \quad \theta_{[k]}^o = \theta_{0,[k]} + \delta$$

with

$$\|\delta\|_2^2 \leq C_1 \|\theta_0 - \theta_{0,[k]}\|_2^2, \quad \|\delta\|_1 \leq C_1 k \frac{\sqrt{\log n}}{\sqrt{n}}$$

as soon as  $k \in \mathcal{K}_n(M)$ , where  $C_1$  depends on  $M$ ,  $\|\theta_0\|_1$  and  $\|\theta_0\|_2$ .

PROOF. The inequalities in (52) are a straightforward consequence of

$$h^2(f_{\theta_2}, f_{\theta_1}) \leq \|f_{\theta_1}\|_\infty^2 \int \left( e^{[(\theta_2 - \theta_1)^T \Phi(x) - c(\theta_2) + c(\theta_1)]/2} - 1 \right)^2 dx$$

and of  $e^{-c_2\|\theta_1\|_1} \leq \|f_{\theta_1}\|_\infty \leq e^{c_1\|\theta_1\|_1}$ .

Next for convenience we introduce the notations  $F_0 = E_{f_0}$  and  $F_\theta = E_{f_\theta}$ . Then by definition,  $\theta_{[k]}^o$  satisfies

$$(53) \quad \frac{\partial \int f_0(x) \log f_\theta(x) dx}{\partial \theta_j} \Big|_{\theta=\theta_{[k]}^o} = -F_0(\phi_j) + F_{\theta_{[k]}^o}(\phi_j) = 0.$$

Write  $\theta_{[k]}^o = \theta_{0,[k]} + \delta$  with  $\theta_{0,[k]} = (\theta_{0,1}, \dots, \theta_{0,k})$  where  $\delta \in \mathbb{R}^k$  and  $\Delta(x) = \sum_{j \leq k} \delta_j \phi_j(x)$ . Write also  $L = \sum_{j > k} \theta_{0,j} \phi_j$  then for all  $j \leq k$ , we have in view of (53)

$$(54) \quad \begin{aligned} & \frac{F_{\theta_{0,[k]}}(\phi_j) + F_{\theta_{0,[k]}}(\phi_j L) + F_{\theta_{0,[k]}}(\phi_j L^2)/2 + F_{\theta_{0,[k]}}(\phi_j R_L)}{1 + F_{\theta_{0,[k]}}(L) + F_{\theta_{0,[k]}}(L^2)/2 + F_{\theta_{0,[k]}}(R_L)} \\ &= \frac{F_{\theta_{0,[k]}}(\phi_j) + F_{\theta_{0,[k]}}(\phi_j \Delta) + F_{\theta_{0,[k]}}(\phi_j \Delta^2)/2 + F_{\theta_{0,[k]}}(\phi_j R_\Delta)}{1 + F_{\theta_{0,[k]}}(\Delta) + F_{\theta_{0,[k]}}(\Delta^2)/2 + F_{\theta_{0,[k]}}(R_\Delta)} \end{aligned}$$

with  $R_L$  (resp.  $R_\Delta$ ) is the error term in the Taylor expansion  $e^L = 1 + L + L^2/2 + R_L$ . Note that in view of (51)  $\|L\|_2 = \sum_{j>k} \theta_{0,j}^2 \lesssim b(k) \lesssim k \log n/n$  for all  $k \in \mathcal{K}_n(M)$  and that  $\|L\|_\infty \leq \|\phi\|_\infty \sum_{j>k} |\theta_{0,j}| = o(1)$  on  $\mathcal{K}_n(M)$ . We also have that

$$F_{\theta_{0,[k]}}(\phi_j \Delta) - F_{\theta_{0,[k]}}(\phi_j) F_{\theta_{0,[k]}}(\Delta) = (\Gamma \delta)_j$$

where  $\Gamma(j_1, j_2) = F_{\theta_{0,[k]}}(\tilde{\phi}_{j_1} \tilde{\phi}_{j_2})$ , and  $\tilde{\phi}_j = \phi_j - F_{\theta_{0,[k]}}(\phi_j)$ . By the assumption  $\int \phi_j = 0$  we have for all  $u \in \mathbb{R}^k$ ,

$$u^T \Gamma u = F_{\theta_{0,[k]}} \left( \left( \sum_{j \leq k} u_j \tilde{\phi}_j \right)^2 \right) \geq c_0 \left\| \sum_{j \leq k} u_j \tilde{\phi}_j \right\|_2^2 = c_0 \|u\|_2^2 + \left( \sum_{j=1}^k F_{\theta_{0,[k]}}(\phi_j) u_j \right)^2 \geq c_0 \|u\|_2^2$$

where  $c_0 \geq e^{-2\|\theta_0\|_1 \|\phi\|_\infty}$ . Therefore (55) can be re-written as

$$\begin{aligned} (55) \quad & (\Gamma \delta)_j + \frac{1}{2} F_{\theta_{0,[k]}}(\tilde{\phi}_j \Delta^2) - F_{\theta_{0,[k]}}(\tilde{\phi}_j \Delta) F_{\theta_{0,[k]}}(\Delta) + \bar{R}_\Delta \\ & = F_{\theta_{0,[k]}}(\tilde{\phi}_j L) + \frac{1}{2} F_{\theta_{0,[k]}}(\tilde{\phi}_j L^2) - F_{\theta_{0,[k]}}(\tilde{\phi}_j L) F_{\theta_{0,[k]}}(L) + \bar{R}_L \end{aligned}$$

where  $\bar{R}_\Delta \lesssim F_0(|\delta|^3)$  and  $\bar{R}_L \lesssim F_0(|L|^3)$ , so that writing  $\tilde{\Phi} = (\tilde{\phi}_1, \dots, \tilde{\phi}_k)^T$  and  $\Phi = (\phi_1, \dots, \phi_k)^T$

$$\begin{aligned} \delta &= \Gamma^{-1} F_{\theta_{0,[k]}}(\tilde{\phi} L) + \frac{1}{2} \Gamma^{-1} F_{\theta_{0,[k]}}(\tilde{\phi} L^2) - \Gamma^{-1} F_{\theta_{0,[k]}}(\tilde{\phi} L) F_{\theta_{0,[k]}}(L) - \frac{1}{2} \Gamma^{-1} F_{\theta_{0,[k]}}(\tilde{\phi} (\Phi^T \Gamma^{-1} F_{\theta_{0,[k]}}(\tilde{\phi} L))^2) \\ &\quad + \Gamma^{-1} F_{\theta_{0,[k]}}(\tilde{\phi} (\Phi^T \Gamma^{-1} F_{\theta_{0,[k]}}(\tilde{\phi} L))) F_{\theta_{0,[k]}}(\Phi)^T \Gamma^{-1} F_{\theta_{0,[k]}}(\tilde{\phi} L)) + \bar{R} \end{aligned}$$

where  $|\bar{R}| \leq \|\delta\|_2^2 \|\delta\|_1$ , moreover

$$\|F_{\theta_{0,[k]}}(\tilde{\phi} L)\|_2^2 = \sum_{j=1}^k \left( (F_{\theta_{0,[k]}}(\tilde{\phi}_j L))^2 \right) \leq 2(\|f_{\theta_{0,[k]}} L\|_2^2 + \|f_{\theta_{0,[k]}}\|_2^2 \|L\|_2^2) \leq 4e^{4\|\theta_0\|_1 \|\phi\|_\infty} \sum_{l>k} \theta_{0,l}^2.$$

So in view of the preceding two displays and (55) there exists a constant  $C_1$  depending on  $\|\theta_0\|_1$  and  $\|\theta_0\|_2$  for which

$$\|\delta\|_2 \leq C_1 \left( \sum_{l>k} \theta_{0,l}^2 \right)^{1/2}$$

Applying (52) to  $\theta_{0,[k]}$  implies  $C^{-1} \|\theta_0 - \theta_{0,[k]}\|_2 \leq h(\theta_0, \theta_{0,[k]}) \leq C \|\theta_0 - \theta_{0,[k]}\|_2$  for some  $C > 0$  so that  $b(k) \lesssim \|\theta_0 - \theta_{0,[k]}\|_2^2$ . Moreover Lemma 3.1 of [34] implies that for  $k \in \mathcal{K}_n(M)$   $\|\theta_0 - \tilde{\theta}\|_2^2 \lesssim b(k) + k \log n/n$  for all  $\tilde{\theta} \in \Theta(k)$  satisfying  $h^2(\theta_0, \tilde{\theta}) \leq b(k) + k \log n/n$ . Hence

$$\|\theta_0 - \theta_{0,[k]}\|_2^2 \leq \|\theta_0 - \tilde{\theta}\|_2^2 \lesssim k \log n/n$$

and  $\|\delta\|_1 \lesssim k \sqrt{\log n} / \sqrt{n} = o(1)$ .  $\square$

**A.4. Proof of Proposition 4.** The proof of the proposition consists of verifying the conditions of Theorem 1. By slightly abusing our notations we write  $h_n^2(\theta, \theta_0) = h_n^2(q_\theta, q_{\theta_0})$  and take  $d(\theta, \theta_0) = h_n(\theta, \theta_0)$  to be the empirical hellinger distance.

First we note that condition **A4** is verified in Lemma 6. Then, similarly to the density example by choosing

$$\Theta_n(k) = \{\theta \in \mathbb{R}^k; \|\theta\|_2 \leq R_n\}, \quad R_n = R_0(n\varepsilon_n(k_n)^2)^{1/p},$$

for some large enough  $R_0 > 0$  condition **A3(k)** holds.

Next we prove condition **A5(k)** for  $k \leq Ak_n \log n$  where  $A$  is given in Theorem 1. Recall from Section 3.4 that  $h_n^2(\theta_1, \theta_2) \leq d_n^2(f_{\theta_1}, f_{\theta_2})$  for any  $\theta_1, \theta_2 \in \Theta$  so that the Hellinger entropy is bounded by the  $d_n$  entropy. Moreover for all  $k \leq K_n$  and  $\theta_1, \theta_2 \in \Theta(k)$ ,

$$d_n^2(f_{\theta_1}, f_{\theta_2}) = (\theta_1 - \theta_2)^T \Phi_k^T \Phi_k (\theta_1 - \theta_2) \leq C_0 \|\theta_1 - \theta_2\|_2^2$$

hence the Hellinger entropy is bounded by the  $\ell_2$  entropy on  $\Theta_n(k)$ . Let  $u > \sqrt{b(k)} \vee u_0 \sqrt{k \log n/n}$  and  $\zeta > 0$ , the covering number of  $\Theta_n(k)$  by  $\ell_2$  balls of radius  $\zeta u$  is bounded from above by a term of order  $\exp(A_1 k (\log n - \log u))$  so that as soon as  $u \geq u_0 \sqrt{k \log n/n} \vee \sqrt{b(k)}$ , the local entropy is bounded by  $A'_1 k \log n$  for positive constant  $A'_1$  and since  $nu^2 > u_0^2 k \log n$ , choosing  $u_0$  large enough, we prove **A5(k)** for all  $k \leq Ak_n \log n$ .

We now study assumption **A6(k)** for all  $(M^2/2)k_n < k \leq Ak_n \log n$ . Let  $\theta$  satisfy  $h_n(\theta_0, \theta)^2 \leq J_0 \varepsilon_n^2(k_n)$ . We shall first prove that  $\max_i |f_\theta(x_i)| \leq M_2$  for some positive constant  $M_2$ . Assume that  $\max_i |f_\theta(x_i)| > M_2$  and split  $\{1, \dots, n\}$  into  $I_1 = \{i; |f_\theta(x_i)| \leq M_2\}$ ,  $I_2 = \{i; f_\theta(x_i) > M_2\}$  and  $I_3 = \{i; f_\theta(x_i) < -M_2\}$  and introduce the notation  $\theta_{0,[k]} = (\theta_{0,1}, \theta_{0,2}, \dots, \theta_{0,k})$ . Then we have for all  $i \in I_2$  and  $l \in \mathbb{N}$  that  $\mu(f_\theta(x_i)) \geq (1+\delta)\mu(f_{\theta_{0,[l]}}(x_i))$  for some  $\delta > 0$  fixed, by choosing  $M_1$  large enough and if  $i \in I_3$ ,  $1 - \mu(f_\theta(x_i)) \geq (1+\delta)(1 - \mu(f_{\theta_{0,[l]}}(x_i)))$ . Moreover for all  $\theta_1, \theta_2 \in \Theta(k)$

$$(56) \quad h_b(f_{\theta_1}(x_i), f_{\theta_2}(x_i))^2 = (f_{\theta_1}(x_i) - f_{\theta_2}(x_i))^2 \frac{\mu'(\bar{f}(x_i))^2}{4\mu(\bar{f}(x_i))(1 - \mu(\bar{f}(x_i)))},$$

for some  $\bar{f}(x_i) \in [f_{\theta_1}(x_i), f_{\theta_2}(x_i)]$ . This implies in particular that for all  $M_1 > 0$  there exist  $c_1, C_1 > 0$  such that if  $\|\theta\|_1 \leq M_1$  and  $\|\theta'\|_1 \leq M_1$  and  $\|\theta - \theta'\|_2 \leq M_1$ , then

$$(57) \quad c_1 d_n(f_\theta, f_{\theta'}) \leq h_n(\theta, \theta') \leq C_1 d_n(f_\theta, f_{\theta'}).$$

Hence, if  $\theta, \theta' \in \Theta(k)$  with  $k \leq K_n$ , then (57) remains valid with  $\|\theta - \theta'\|_2$  replacing  $d_n(f_\theta, f_{\theta'})$ .

Let  $k_n^* = k_0(n/\log n)^{1/(2\beta+1)} \vee Ak_n \log n$  so that in view of (42)

$$(58) \quad h_n(\theta_0, \theta_{0,[k_n^*]}) \lesssim d_n(f_{\theta_0}, f_{\theta_{0,[k_n^*]}}) \lesssim (k_n^*)^{-\beta}.$$

Furthermore, in view of  $\|f_{\theta_{0,[k_n^*]}}\|_\infty \leq \max_j \|\phi_j\|_\infty \|\theta_0\|_1$  we obtain that if  $|f_\theta(x_i)| \leq M_1$  then there exists  $c_1 > 0$  such that  $h_b^2(f_\theta(x_i), f_{\theta_{0,[k_n^*]}}(x_i)) \geq c_1(f_\theta(x_i) - f_{\theta_{0,[k_n^*]}}(x_i))^2$ . This implies in particular that

$$\begin{aligned} nh_n^2(\theta, \theta_{0,[k_n^*]}) &\geq c_1 \sum_{i \in I_1} (f_\theta(x_i) - f_{\theta_{0,[k_n^*]}}(x_i))^2 + \delta \sum_{i \in I_2 \cup I_3} \mu(f_{\theta_{0,[k_n^*]}}(x_i)) \wedge (1 - \mu(f_{\theta_{0,[k_n^*]}}(x_i))) \\ &\geq c_1 \sum_{i \in I_1} (f_\theta(x_i) - f_{\theta_{0,[k_n^*]}}(x_i))^2 + \delta c_0 |I_2 \cup I_3| \\ &\geq c_1 (\theta - \theta_{0,[k_n^*]})^T \Phi_{I_1}^T \Phi_{I_1} (\theta - \theta_{0,[k_n^*]}) + \delta c_0 |I_2 \cup I_3| \end{aligned}$$

where  $\Phi_{I_1}$  whose rows are given by  $(\phi_1(x_i), \dots, \phi_{k_n^* \vee k}(x_i))$  for all  $i \in I_1$ . We also note that in view assertion (42) and  $b(k) \leq \tilde{b}(k)$  we get  $b(k) \lesssim k^{-2\beta}$ . Thus

$$\Phi_{I_1}^T \Phi_{I_1} = \Phi_{k_n^* \vee k}^T \Phi_{k_n^* \vee k} - \Phi_{I_1^c}^T \Phi_{I_1^c}$$

and for all  $j_1, j_2 \in \{1, \dots, k_n^* \vee k\}$ , take  $\tilde{\theta} \in \Theta(k)$  such that  $h^2(\tilde{\theta}, \theta_0) \leq \inf_{\theta \in \Theta(k)} h^2(\theta, \theta_0) + (n/\log n)^{-2\beta/(2\beta+1)}$

$$\begin{aligned} |(\Phi_{I_1^c}^T \Phi_{I_1^c})(j_1, j_2)| &\leq \max_j \|\phi_j\|_\infty^2 |I_2 \cup I_3| \lesssim nh_n^2(\tilde{\theta}, \theta_{0,[k_n^*]}) \\ &\lesssim nb(k) + nh_n^2(\theta_0, \theta_{0,[k_n^*]}) \lesssim nk^{-2\beta} + (n/\log n)^{-2\beta/(2\beta+1)} n. \end{aligned}$$

From the computation in Section 3.4,  $k_n \lesssim (n/\log n)^{1/(2\beta+1)}$  so that

$$|(\Phi_{I_1^c}^T \Phi_{I_1^c})(j_1, j_2)|/n \lesssim k^{-2\beta} + (n/\log n)^{-2\beta/(2\beta+1)} = o(1/k)$$

for all  $k = o(n/\log n)^{2\beta/(2\beta+1)}$  which is the case when  $k \lesssim Ak_n \log n$  and  $\beta \geq \beta_0 > 1/2$ . Hence

$$(59) \quad \frac{\Phi_{I_1}^T \Phi_{I_1}}{n} \asymp I_d$$

and in view of the computations in Section 3.4

$$\begin{aligned} \|\theta - \theta_{0,[k_n^*]}\|_2^2 &\lesssim \frac{1}{n} (\theta - \theta_{0,[k_n^*]})^T \Phi_{I_1}^T \Phi_{I_1} (\theta - \theta_{0,[k_n^*]}) \\ &\lesssim h_n^2(\theta, \theta_{0,[k_n^*]}) \leq h_n^2(\theta, \theta_0) + h_n^2(\theta_0, \theta_{0,[k_n^*]}) \\ &\lesssim \varepsilon_n^2(k_n) + (k_n^*)^{-2\beta} \lesssim (n/\log n)^{-2\beta/(2\beta+1)} \log n. \end{aligned}$$



This implies in particular that  $\|\theta - \theta_{0,[k_n^*]}\|_1 \lesssim \sqrt{k}(n/\log n)^{-\beta/(2\beta+1)}\sqrt{\log n} = o(1)$  for  $k \lesssim Ak_n \log n$ . Hence  $\|f_\theta\|_\infty \leq \|f_0\|_\infty + o(1)$  for  $n$  large enough.

Now let  $\theta \in B_k(\theta_0, J_0\varepsilon_n(k))$ , then  $\theta \in B_k(\theta_{0,[k_n^*]}, J_0\varepsilon_n(k) + C(k_n^*)^{-\beta})$  for some  $C > 0$  and there exist  $C_1, C_2, C_3, C_4, C_5 > 0$  constants such that

$$\begin{aligned} \text{Vol}(B_k(\theta_0, J_0\varepsilon_n(k))) &\leq \text{Vol}(B_k(\theta_{0,[k_n^*]}, C_1(J_0\varepsilon_n(k) + C(k_n^*)^{-\beta}), \|\cdot\|_2)) \\ &\leq e^{C_2 k \log(C_3[\varepsilon_n(k) + (k_n^*)^{-\beta}])} \\ &\leq e^{C_2 k \log(C_4(k(\log n)/n + (n/\log n)^{-\beta/(1+2\beta)}))} \\ &\leq e^{-C_5(M^2/2)k_n \log n}, \end{aligned}$$

which, similarly to the proof of Proposition 3 is bounded from above by  $e^{-C_6 n \varepsilon_n^2(k_n)}$  for arbitrary large constant  $C_6$  by appropriate choice of the parameter  $M > 0$ , providing us **A6(k)** for all  $k_n M^2/2 \leq k \leq Ak_n \log n$ .

Next we deal with condition **A2(k)** for  $k \in \mathcal{K}_n(M)$ . Let  $\theta_{[k]}^o$  be the Kullback-Leibler projection of  $\theta_0$  onto  $\Theta(k)$ ,  $\theta_{[k]}^o = \theta_{0,[k]} + \delta$ . From Lemma 5, since  $K_n \gg n^{\frac{1}{2(\beta_0-1/2)}}$ , then  $\sqrt{k}K_n^{-(\beta_0-1/2)} = o(\sqrt{k/n})$  so that

$$\|\delta\|_2 \leq C_1 \|\theta_0 - \theta_{0,[k]}\|_2 + o(\sqrt{k/n}), \quad \|\delta\|_2 = o(1).$$

This implies in particular that

(60)

$$\begin{aligned} p_{\theta_0}^{(n)} \log \frac{p_{\theta_{[k]}^o}^{(n)}}{p_\theta^{(n)}} &= \sum_i q_0(x_i) (\theta_{[k]}^o - \theta)^T \Phi_k(x_i) + \log(1 + e^{\theta^T \Phi_k(x_i)}) - \log(1 + e^{(\theta_{[k]}^o)^T \Phi_k(x_i)}) \\ &= \sum_i (q_0(x_i) - q_{\theta_{[k]}^o}(x_i)) (\theta_{[k]}^o - \theta)^T \Phi_k(x_i) + O\left(\sum_i \left((\theta_{[k]}^o - \theta)^T \Phi_k(x_i)\right)^2\right) \\ &= O\left(\sum_i \left((\theta_{[k]}^o - \theta)^T \Phi_k(x_i)\right)^2\right) = O(n \|\theta_{[k]}^o - \theta\|_2^2) = O(n h_n^2(\theta_{[k]}^o, \theta)), \end{aligned}$$

where in the second line we used the Taylor expansions of  $f(\theta) = \log(1 + e^{\theta^T \Phi_k(x_i)})$  around  $\theta_{[k]}^o$ , while the third line follows from Lemma 5. Similarly we obtain

$$p_{\theta_0}^{(n)} \log^2 \left( \frac{p_{\theta_{[k]}^o}^{(n)}}{p_\theta^{(n)}} \right) \lesssim n \|\theta_{[k]}^o - \theta\|_2^2 = O(n h_n^2(\theta_{[k]}^o, \theta))$$

so that first assertion of **A2(k)** is verified. The second assertion then follows from (60) along the same lines of reasoning as in the proof of Proposition 1.

Similarly, since for  $k \in \mathcal{K}_n(M)$  and  $\tilde{\theta} \in B_k(\theta_{[k]}^o, M_0 \varepsilon_n(k_n))$ ,  $\|\tilde{\theta}\|_1 \leq \|\theta_{[k]}^o\|_1 + o(1) = \|\theta_0\|_1 + o(1)$ , for all  $\theta \in B_k(\tilde{\theta}, \delta \sqrt{k/n})$  we have  $h_n(\tilde{\theta}, \theta) \gtrsim \|\tilde{\theta} - \theta\|_2$ , then

$$\text{Vol}(B_k(\tilde{\theta}, \delta \sqrt{k/n})) \leq \text{Vol}(B_k(\tilde{\theta}, C_1 \delta \sqrt{k/n}), \|\cdot\|_2)$$

and from (57), there exists  $1 \geq \rho > 0$  such that if  $\|\theta - \theta_{[k]}^o\|_2 \leq \rho \sqrt{k/n}$  then  $h_n(\theta_{[k]}^o, \theta) \leq \sqrt{k/n}$  which in turns implies that

$$\text{Vol}(B_k(\theta_{[k]}^o, \sqrt{k/n})) \geq \text{Vol}(B(\theta_{[k]}^o, \rho \sqrt{k/n}, \|\cdot\|_2))$$

which proves **A7(k)**.

Finally we deal with condition **A1**. Following from (58) we have  $h(\theta, \theta_{0,[k_n^*]}) \leq h(\theta, \theta_0) + h(\theta_0, \theta_{0,[k_n^*]}) \lesssim \varepsilon_n(k_n) + (k_n^*)^{-\beta}$ . Hence in view of (59) we get

$$\|\theta - \theta_{0,[k_n^*]}\|_1 \leq \sqrt{k_n} \|\theta - \theta_{0,[k_n^*]}\|_2 \lesssim \sqrt{k_n} h(\theta, \theta_{0,[k_n^*]}) = o(1)$$

and as a direct consequence  $\|f_\theta\| = O(1)$ . Therefore, in view of (56) and

$$\sum_{i=1}^n (\log q_{\theta_0}(x_i) - \log q_\theta(x_i))^2 = \sum_{i=1}^n \frac{\mu'(\bar{f}(x_i))}{\mu(\bar{f}(x_i))} (f_{\theta_0}(x_i) - f_\theta(x_i))^2,$$

for some  $\bar{f}(x_i) \in [f_\theta(x_i), f_{\theta_0}(x_i)]$ ,  $i = 1, \dots, n$ , (with  $\|\bar{f}\|_\infty < \infty$ ) we get that  $KL(\theta_0, \theta) \lesssim h_n^2(\theta_0, \theta)$ . Similarly we can also get that  $V(\theta_0, \theta) \lesssim h_n^2(\theta_0, \theta)$  provinding us the first assertion of **A1**. To prove the second part of the condition take any  $\tilde{\theta} \in \Theta(k_n)$  satisfying  $h_n(\tilde{\theta}, \theta_0) \leq \inf_{\theta \in \Theta(k)} h_n(\theta, \theta_0) + k_n \log n / (2n)$ . Then similarly to the proof of condition **A2(k)** there exists  $c > 0$  such that

$$B_{k_n}(\theta_0, \varepsilon_n(k_n)) \supset B_{k_n}(\tilde{\theta}, k_n \log n / (2n)) \supset B_{k_n}(\tilde{\theta}, ck_n \log n / (2n), \|\cdot\|_2)$$

and the prior probability of this small ball can bounded from below using the standard computations as in the proof of Proposition 1.

It remained to show the second statement of the lemma. Again as a consequence of Corollary 1 it is sufficient to verify that  $\varepsilon_n(k_n) \lesssim (n / \log n)^{-\beta / (1+2\beta)}$ , which follows automatically from the computations in Section 3.4, where the bound  $k_n \lesssim (n / \log n)^{1/(1+2\beta)}$  was derived.

**LEMMA 5.** *Let  $\theta_{[k]}^o$  be the Kullback-Leibler projection of  $\theta_0$  onto  $\Theta(k)$ . If  $\theta_0 \in \mathcal{S}^{\beta_0}(M_0)$  for some  $M_0 > 0$  and  $\beta_0 > 1/2$ , then  $\theta_{[k]}^o$  satisfies*

$$(61) \quad \forall j \leq k \quad \sum_{i=1}^n (q_0(x_i) - q_{\theta_{[k]}^o}(x_i)) \phi_j(x_i) = 0, \quad q_{\theta_{[k]}^o}(x) = \frac{e^{(\theta_{[k]}^o)^T \Phi_k(x)}}{1 + e^{(\theta_{[k]}^o)^T \Phi_k(x)}}$$

and writing  $\theta_{[k]}^o = \theta_{0,[k]} + \delta$ , then there exist  $C_1 > 0$  depending on  $\|\theta_0\|_1$  and  $\|\theta_0\|_2$  such that for all  $k \in \mathcal{K}_n(M)$

$$\|\delta\|_2 \leq C_1 \|\theta_0 - \theta_{0,[k]}\|_2 + \sqrt{k} K_n^{-\beta_0+1/2}, \quad \|\delta\|_1 \leq \frac{k\sqrt{\log n}}{\sqrt{n}} + k K_n^{-\beta_0+1/2}$$

PROOF OF LEMMA 5. The proof is very similar to the proof of Lemma 4. Equality (61) is a direct consequence of the definition of  $\theta_{[k]}^o$ . Moreover writing  $\sum_{j=1}^{\infty} \theta_{0,j} \phi_j(x) = \theta_{0,[k]}^T \Phi_k(x) + L$ , where  $L(x) = \sum_{j=k+1}^{\infty} \theta_{0,j} \phi_j(x)$ , and  $\bar{L} = e^L - 1$

$$\begin{aligned} q_0(x) &= q_{0,[k]}(x) \frac{e^{L(x)}}{1 + q_{0,[k]}(x)(e^{L(x)} - 1)} = q_{0,[k]}(x) + \frac{\bar{L}(x) q_{0,[k]}(x)(1 - q_{0,[k]}(x))}{1 + q_{0,[k]}(x)\bar{L}(x)} \\ &= q_{0,[k]}(x) + \bar{L}(x) q_{0,[k]}(x)(1 - q_{0,[k]}(x)) + O(\bar{L}(x)^2) \end{aligned}$$

and as in Lemma 4,  $\|\bar{L}\|_2^2 \lesssim \|L\|_2^2 \lesssim k \log n/n$  and  $\|\bar{L}\|_{\infty} \lesssim \|L\|_{\infty} = o(1)$ . For all  $\|\delta\|_1 \leq M_0$ ,  $\delta \in \mathbb{R}^k$  with some given  $M_0$  and  $\|\delta\|_2$  small enough, then the same expansion of  $q_{\theta_{0,[k]}+\delta}$  leads to

$$q_{\theta_{0,[k]}+\delta} = q_{0,[k]}(x) + \delta^T \Phi_k(x) q_{0,[k]}(x)(1 - q_{0,[k]}(x)) + O(|\delta^T \Phi_k(x)|^2).$$

Moreover let  $z \in \mathbb{R}^k$ , then

$$\begin{aligned} &\sum_{j_1, j_2=1}^k z_{j_1} z_{j_2} \sum_{i=1}^n q_{0,[k]}(x_i)(1 - q_{0,[k]}(x_i)) \phi_{j_1}(x_i) \phi_{j_2}(x_i) \\ &= \sum_{i=1}^n q_{0,[k]}(x_i)(1 - q_{0,[k]}(x_i)) \left( \sum_{j=1}^k z_j \phi_j(x_i) \right)^2 \\ &\geq \min_x q_{0,[k]}(x)(1 - q_{0,[k]}(x)) z^T \Phi_k^T \Phi_k z \\ &\gtrsim \|z\|_2^2 \end{aligned}$$

by assumption on the matrix  $\Phi_k$ . Define the matrix  $\Gamma_n$  with  $j_1, j_2$  coefficients  $n^{-1} \sum_{i=1}^n q_{0,[k]}(x_i)(1 - q_{0,[k]}(x_i)) \phi_{j_1}(x_i) \phi_{j_2}(x_i)$ ,  $j_1, j_2 \leq K_n$ . Then  $\Gamma_n$  is equivalent to the identity matrix on  $\mathbb{R}^{K_n}$ . Denote  $\Gamma_{n,[k]}$  the upper square sub-matrix of  $\Gamma_n$  of dimension  $k$ . Combined with (61), this implies in particular that

$$\delta = \Gamma_{n,[k]}^{-1} \Delta_n, \quad \Delta_n(j) = \frac{1}{n} \sum_{i=1}^n L(x_i) \phi_j(x_i) q_{0,[k]}(x_i)(1 - q_{0,[k]}(x_i)) + O\left(\sum_{i=1}^n L(x_i)^2/n\right)$$

We then have, using the fact that  $\|L\|_\infty = o(1)$

$$\begin{aligned} \|\delta\|_2^2 &\lesssim \|\Delta_n\|_2^2 \lesssim \sum_{j=1}^k \left( \frac{1}{n} \sum_{i=1}^n L(x_i) \phi_j(x_i) q_{0,[k]}(x_i) (1 - q_{0,[k]}(x_i)) \right)^2 \\ &\quad + O \left( k \left( \sum_{i=1}^n L(x_i)^2 / n \right)^2 \right). \end{aligned}$$

We have that

$$\frac{\sum_{i=1}^n L(x_i)^2}{n} = d_n(f_0, f_{\theta_{0,[k]}})^2 \lesssim K_n^{-2(\beta_0-1/2)} + \|L\|_2^2 \lesssim K_n^{-2(\beta_0+1/2)} + k \log n / n.$$

Then writing  $\Gamma_{n[>k,j]}$  the vector with  $l$  coefficient ( $K_n > l > k$ )  $\Gamma_n(l, j)$  and  $\Gamma_{n,[>k],[k]}$  the matrix whose  $j$ -th column is  $\Gamma_{n[>k,j]}$ ,  $j \leq k$

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n L(x_i) \Phi_j(x_i) q_{0,[k]}(x_i) (1 - q_{0,[k]}(x_i)) \\ &= \theta_{0,[k:K_n]}^T \Gamma_{n[>k,j]} + \sum_{l>K_n} \theta_{0,l} \frac{1}{n} \sum_{i=1}^n \phi_l(x_i) \phi_j(x_i) q_{0,[k]}(x_i) (1 - q_{0,[k]}(x_i)) \\ &= \theta_{0,[k:K_n]}^T \Gamma_{n[>k,j]} + O(K_n^{-\beta_0+1/2}) \end{aligned}$$

if  $f_0 \in \mathcal{S}^{\beta_0}(M_0)$ , where  $\theta_{0,[k:K_n]} = (\theta_{0,k+1}, \dots, \theta_{0,K_n})$ . Since

$$\sum_{j \leq k} (\theta_{0,[k:K_n]}^T \Gamma_{n[>k,j]})^2 \leq \|\Gamma_n \begin{pmatrix} 0_{[k]} \\ \theta_{0,[k:K_n]} \end{pmatrix}\|_2^2 \lesssim \|L\|_2^2$$

we obtain that

$$\|\delta\|_2^2 \lesssim \|L\|_2^2 + k K_n^{-2\beta_0+1}.$$

□

LEMMA 6. *In the classification model (26) for all  $M > 0$  there exists a  $B > 0$  such that*

$$P_{\theta_0}^{(n)} \left( \sup_{k \in \mathcal{K}_n} \sup_{\theta \in \Theta_n(k) \cap B_k(\theta_{[k]}^o, M \rho_n \varepsilon_n(k))} \ell_n(\theta) - \ell_n(\theta_{[k]}^o) - Bk > 0 \right) = o(1).$$

PROOF. First of all note that by Cauchy-Schwarz we get

$$\begin{aligned}
 & \left| \ell_n(\theta) - \ell_n(\theta_{[k]}^o) - E_{\theta_0}^{(n)}(\ell_n(\theta) - \ell_n(\theta_{[k]}^o)) \right| \\
 &= \left| \sum_{i=1}^n \left( \log q_\theta(x_i) - \log q_{\theta_{[k]}^o}(x_i) \right) (y_i - q_{\theta_0}(x_i)) \right. \\
 &\quad \left. + \sum_{i=1}^n \left( \log [1 - q_\theta(x_i)] - \log [1 - q_{\theta_{[k]}^o}(x_i)] \right) (q_{\theta_0}(x_i) - y_i) \right| \\
 &\leq \sqrt{\sum_{i=1}^n (y_i - q_{\theta_0}(x_i))^2} \times \left\{ \sqrt{\sum_{i=1}^n \left( \log q_\theta(x_i) - \log q_{\theta_{[k]}^o}(x_i) \right)^2} \right. \\
 (62) \quad & \left. + \sqrt{\sum_{i=1}^n \left( \log [1 - q_\theta(x_i)] - \log [1 - q_{\theta_{[k]}^o}(x_i)] \right)^2} \right\}.
 \end{aligned}$$

Then note that

$$(63) \quad E_{\theta_0}^{(n)} \sum_{i=1}^n (y_i - q_{\theta_0}(x_i))^2 \leq n.$$

Furthermore, in view of Lemma 5, we have for  $\theta \in B_k(\theta_{[k]}^o, M\rho_n\varepsilon_n(k))$ ,  $k \in \mathcal{K}_n(M)$  that  $\|\theta - \theta_{[k]}^o\|_1 \leq \sqrt{k}\|\theta - \theta_{[k]}^o\|_2 \lesssim \sqrt{k_n}\rho_n\varepsilon_n(k_n) = O(1)$  and therefore  $\|f_\theta\|_\infty = O(1)$ . Hence for some  $\bar{f}(x_i) \in [f_\theta(x_i), f_{\theta_{[k]}^o}(x_i)]$ ,  $i = 1, 2, \dots, n$  (and therefore bounded  $\bar{f}(x_i)$ ) that

$$\sum_{i=1}^n \left( \log q_\theta(x_i) - \log q_{\theta_{[k]}^o}(x_i) \right)^2 \leq \sum_{i=1}^n \frac{\mu'(\bar{f}(x_i))}{\mu(\bar{f}(x_i))} (f_\theta(x_i) - f_{\theta_{[k]}^o}(x_i))^2 \lesssim d_n^2(f_\theta, f_{\theta_{[k]}^o})$$

and similarly

$$\sum_{i=1}^n \left( \log [1 - q_\theta(x_i)] - \log [1 - q_{\theta_{[k]}^o}(x_i)] \right)^2 \lesssim d_n^2(f_\theta, f_{\theta_{[k]}^o}).$$

We conclude the proof by applying markov's inequality and (60).  $\square$

## REFERENCES

- [1] Arbel, J., Gayraud, G., and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics*, 40(3):549–570.
- [2] Belitser, E. (2014). On coverage and local radial rates of DDM-credible sets. *ArXiv e-prints*.

- [3] Belitser, E. and Nurushev, N. (2015). Needles and straw in a haystack: empirical bayes confidence for possibly sparse sequences. *arXiv preprint arXiv:1511.01803*.
- [4] Birge, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65:181–237.
- [5] Birgé, L. (1983). Robust testing for independent non identically distributed variables and markov chains. In *Specifying Statistical Models*, volume 16 of *Lecture Notes in Statistics*, pages 134–162. Springer New York.
- [6] Bull, A. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Statist.*, 6:1490–1516.
- [7] Bull, A. D. and Nickl, R. (2013). Adaptive confidence sets in  $l^2$ . *Probability Theory and Related Fields*, 156(3-4):889–919.
- [8] Cai, T. and Low, M. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.*, 32:1805–1840.
- [9] Carpentier, A. (2013). Honest and adaptive confidence sets in  $l_p$ . *Electron. J. Statist.*, 7:2875–2923.
- [10] Carpentier, A. and Nickl, R. (2015). On signal detection and confidence sets for low rank inference problems. *Electron. J. Statist.*, 9(2):2675–2688.
- [11] Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein-von Mises theorems in gaussian white noise. *Ann. Statist.*, 41(4):1999–2028.
- [12] Castillo, I. and Nickl, R. (2014). On the Bernstein-von Mises phenomenon for nonparametric bayes procedures. *Ann. Statist.*, 42(5):1941–1969.
- [13] Castillo, I. and Rousseau, J. (2015). A Bernstein-von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.*, 43(6):2353–2383.
- [14] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.*, 42(5):1787–1818.
- [15] Cox, D. D. (1993). An analysis of bayesian inference for nonparametric regression. *Ann. Statist.*, 21(2):903–923.
- [16] Freedman, D. (1999). On the Bernstein Von Mises theorem with infinite dimensional parameter. *Ann. Statist.*, 27:1119–1140.
- [17] Ghosal, S., Ghosh, J. K., and van der Vaart, A. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28:500–531.
- [18] Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non iid observations. *Ann. Statist.*, 35(1):192–223.
- [19] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.*, 38(2):1122–1170.
- [20] Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge series in statistical and probabilistic mathematics.
- [21] Hoffmann, M. and Nickl, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.*, 39:2383–2409.
- [22] Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2013). On adaptive posterior concentration. Technical report.
- [23] Kirichenko, A. and van Zanten, H. (2015). Estimating a smooth function on a large graph by Bayesian Laplacian regularisation. *ArXiv e-prints*.
- [24] Knapik, B., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with gaussian priors. *Ann. Statist.*, 39(5):2626–2657.
- [25] Leahu, H. (2011). On the bernstein-von mises phenomenon in the Gaussian white noise model. *Electronic journal of statistics*, 5:373–405.
- [26] Low, M. (1997). On nonparametric confidence intervals. *Ann. Statist.*, 25:2547–2554.
- [27] Nickl, R. and Szabó, B. (2014). A sharp adaptive confidence ball for self-similar functions. *ArXiv e-prints*.

- [28] Nickl, R. and van de Geer, S. (2013). Confidence sets in sparse regression. *Ann. Statist.*, 41(6):2852–2876.
- [29] Petrone, S., Rousseau, J., and Scricciolo, C. (2014). Bayes and empirical Bayes: do they merge? *Biometrika*, 101:285–302.
- [30] Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.*, 28(1):298–335.
- [31] R. McVinish, J. Rousseau, K. M. (2009). Bayesian goodness-of-fit testing with mixtures of triangular distributions. *Scandinavian Journ. Statist.*, 36:337–354.
- [32] Ray, K. (2014). Adaptive Bernstein-von Mises theorems in Gaussian white noise. *ArXiv e-prints*.
- [33] Rivoirard, V. and Rousseau, J. (2012a). On the Bernstein Von Mises theorem for linear functionals of the density. *Ann. Statist.*, 40:1489–1523.
- [34] Rivoirard, V. and Rousseau, J. (2012b). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis*, 7:311–334.
- [35] Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.*, 34(1):229–253.
- [36] Rousseau, J. and Szabo, B. (2015). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *To appear in Annals of Statistics*.
- [37] Rousseau, J. and Szabo, B. (2015). Asymptotic behaviour of the empirical bayes posteriors associated to maximum marginal likelihood estimator: supplementary material. Technical report.
- [38] Scricciolo, C. (2007). On rates of convergence for bayesian density estimation. *Scandinavian Journal of Statistics*, 34(3):626–642.
- [39] Serra, P. and Krivobokova, T. (2014). Adaptive empirical Bayesian smoothing splines. *ArXiv e-prints*.
- [40] Sniekers, S. and van der Vaart, A. (2015). Adaptive Bayesian credible sets in regression with a Gaussian process prior. *Electron. J. Stat.*, 9(2):2475–2527.
- [41] Söhl, J. and Trabs, M. (2014). Adaptive confidence bands for Markov chains and diffusions: Estimating the invariant measure and the drift. *ArXiv e-prints*.
- [42] Szabó, B. (2015). *Bayesian Statistics from Methods to Models and Applications: Research from BAYSM 2014*, chapter On Bayesian Based Adaptive Confidence Sets for Linear Functionals, pages 91–105. Springer International Publishing, Cham.
- [43] Szabó, B., van der Vaart, A., and van Zanten, H. (2015). Honest bayesian confidence sets for the l2-norm. *Journal of Statistical Planning and Inference*, 166:36 – 51. Special Issue on Bayesian Nonparametrics.
- [44] Szabo, B. T., van der Vaart, A. W., and van Zanten, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Annals of Statistics*, 43(4):1391–1428.
- [45] Tsybakov, A. (2008). *An Introduction to Nonparametric Estimation*. Springer-Verlag, New York.
- [46] van der Pas, S., Szabo, B., and van der Vaart, A. (2016). How many needles in the haystack? adaptive inference and uncertainty quantification for the horseshoe. *arXiv*.
- [47] van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463.
- [48] Verdinelli, I. and Wasserman, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.*, 26:1215–1241.
- [49] Yoo, W. W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.*, 44(3):1069–1102.

CEREMADE, UNIVERSITY PARIS DAUPHINE  
PLACE DU MARÉCHAL DE LATTRE DE TASSIGNY  
E-MAIL: [rousseau@ceremade.dauphine.fr](mailto:rousseau@ceremade.dauphine.fr)

LEIDEN UNIVERSITY,  
MATHEMATICAL INSTITUTE,  
NIELS BOHRWEG 1, LEIDEN, 2333 CA,  
THE NETHERLANDS  
E-MAIL: [b.t.szabo@math.leidenuniv.nl](mailto:b.t.szabo@math.leidenuniv.nl)